REMARKS

Claims 1-2, 7, 9, 12-15, 17-18, 20-21, and 23, drawn to non-elected subject matter, were withdrawn from consideration by the Examiner.

Claims 16, 19, and 22 are canceled.

Claims 3-6, 8, and 10-11 are under consideration.

Claims 3, as currently amended, is in independent form and now contains all of the limitations of original claim 1. Support for this amendment can be found in original claim 1.

Claim 8, as currently amended, depends from claim 3, which is presently amended to contain the limitations of original claim 1. Support for this amendment can be found in original claim 1.

Claims 3, 8, and 10, as currently amended, recite SEQ ID NO:5 or SEQ ID NO:19. Support for these amendments can be found in the corresponding original claims.

Claims 3 and 8 have been amended to recite "wherein said biologically active fragment has sphingosine kinase activity" to further clarify the intended subject matter of the claimed invention. Support for these amendments can be found, for example, in the specification at p. 59.

Claim 3 has been amended to recite "comprising at least 150 contiguous amino acids" and claim 11 has been amended to recite "500 contiguous nucleotides." Support for these amendments can be found, for example, in the specification at p. 13, line 40 through p. 14, line 4.

No new matter has been added by any of these amendments. Entry of these amendments is therefore respectfully requested.

Applicants reserve the right to prosecute non-elected subject matter in subsequent divisional applications.

Restriction Requirement

Applicants respectfully reiterate our traversal to the restriction requirement for at least the reasons already made of record.

Moreover, Applicants note that, as currently amended, all of the claims currently under consideration avoid the prior art cited by the Examiner as destroying unity of invention. They reiterate their request that the Examiner examine at least claims directed to the polypeptide of

SEQ ID NO:5 and the polynucleotide sequence of SEQ ID NO:19 in this single application.

Objections

Claims 3-6 and 8 are objected to as depending from a non-elected claim (claim 1). Claim 3, as currently amended, is in independent form and now contains all of the limitations of original claim 1. Claim 8 now depends from claim 3. Claims 4-6 no longer depend from a non-elected claim. Applicants respectfully request that these amendments be entered and that this objection be withdrawn.

Claims 3-6, 8, and 10 are objected to as containing non-elected subject matter (i.e., SEQ ID NO:1-4, SEQ ID NO:6-18, and SEQ ID NO:20-28). The claims, as currently amended, recite the polypeptide of SEQ ID NO:5 and fragments and variants thereof, and the encoding polynucleotide of SEQ ID NO:19 and fragments and variants thereof. Applicants respectfully request that these amendments be entered and that this objection be withdrawn.

Rejection under 35 U.S.C. §112, 2nd

Claims 3-6 and 8 are rejected under 35 U.S.C. §112, 2nd as allegedly being indefinite due to the recitation of "biologically active" in original claim 1. Claims 3 and 8, as currently amended, recite the limitations of original claim 1. These claims have been further amended to recite "wherein said biologically active fragment has sphingosine kinase activity" in order to further clarify the meaning of the term "biologically active." One of skill in the art would clearly understand that the "biological activity" to which the claims as amended refer is sphingosine kinase activity, as this specific activity is now recited explicitly. These amendments are fully supported by the disclosure in the present application, and are put forth merely to further clarify the claims and to obtain expeditious allowance of the instant application. Applicants expressly do not disclaim equivalents of the invention which could include polypeptides having additional biological activities other than sphingosine kinase inducing activity. Therefore, Applicants respectfully request that the rejection under 35 U.S.C. § 112, second paragraph be withdrawn.

Rejection under 35 U.S.C. § 101 and 35 U.S.C § 112, 1st paragraph

Claims 3-6, 8, and 10-11 stand rejected under 35 U.S.C. §§ 101 and 112, first paragraph, based on the allegation that the claimed invention lacks patentable utility. The rejection alleges in particular that "the specification fails to assert what compounds the protein of SEQ ID NO:5 phosphorylates" and that therefore "the skilled artisan would require further research to identify or reasonable confirm a real world context of use." Applicants traverse this rejection for at least the following reasons.

In response to this issue, Applicants direct the Examiner's attention to those points in the specification that detail the specificity of the claimed polypeptide. In particular, Applicants direct the Examiner's attention to the specification at p. 23, lines 11-18. These lines describe the methods with which the claimed polypeptides are characterized. In particular, "column 5 [of Table 2] shows the amino acid residues comprising signature sequences and motifs; column 6 shows homologous sequences as identified by BLAST analysis" and "[t]he methods of column 7 were used to characterize each polypeptide through sequence homology and protein motifs."

Turning the Examiner's attention to Table 2 (p. 59, last row), mouse sphingosine kinase is cited as a homologous sequence. As the Examiner has recognized, SEQ ID NO:5 is over 80% identical to mouse sphingosine kinase. The Examiner has however, seemingly disregarded this recitation of a homolog and focused on the identification of a diacylglycerol kinase catalytic domain, alleging on that basis that SEQ ID NO:5 is a diacylglycerol kinase while admitting that no such assertion was made by Applicants.

As stated below (Section II.C.), the presence of this domain *supports* the characterization of SEQ ID NO:5 as a sphingosine kinase and indeed, as it was known in the art at the time of filing of the instant application, diacylglycerol kinases and sphingosine kinases share regions of significant homology (Kohama *et al.*, JBC 273:23722-8, 1998). Therefore, the identification of a diacylglycerol kinase catalytic domain is consistent with the identification of mouse sphingosine kinase as a homologous sequence. Thus, there is sufficient basis in the specification for identifying SEQ ID NO:5 as a sphingosine kinase. Indeed, upon reading of the specification and attached tables, the skilled artisan would have had no reason to doubt that the characterization of SEQ ID NO:5 as a human homolog to the disclosed mouse sphingosine kinase.

Furthermore, an alignment of SEQ ID NO:5 with a post-filing human sphingosine kinase (Nava et al., FEBS Letters 473:81-4 (2000); Reference No. 1) shows that the two sequences are approximately 99% identical over the entire 384 amino acid residue length of both sequences.

Thus Applicants' assertion that SEQ ID NO:5 is a sphingosine kinase is corroborated by post-filing experimental data.

In sum, no additional research would be required of the skilled artisan to find a real world context of use for the claimed invention, as its use as a sphingosine kinase is sufficiently asserted in the specification and further supported by additional data in the literature.

As a preliminary matter Applicants respond to an issue the Examiner raised as part of the rejection under 35 U.S.C. § 101. The Examiner alleges that although Applicants assert that the claimed polynucleotides are useful for the diagnosis, treatment or prevention of neurological, cell proliferative and autoimmune/inflammatory disorders, there is no link of SEQ ID NO:19 to a specific disease state (see Office Action at pp. 7-8). Applicants respectfully disagree and, by way of example, direct the Examiner's attention to p. 57 of the specification (Table 1, row 6). Column 5 shows a list of specific cDNA libraries in which fragments of SEQ ID NO:19 were expressed. For example, fragment 1519153H1 was expressed in a cDNA library (BLADTUT04) derived from bladder tumor tissue (for a description of this library, see Table 4, p. 65, row 2). These data support the assertion that SEQ ID NO:19 may be useful in the diagnosis, treatment or prevention of this cell proliferative disorder by showing that SEQ ID NO:19 is expressed in at least one cell proliferative disorder that of bladder cancer.

The rejection of claims 3-6, 8, and 10-11 **is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well known to one of ordinary skill in the art.**

The invention at issue is a polynucleotide corresponding to a gene that is expressed in humans. The novel polynucleotide codes for a polypeptide demonstrated in the patent specification to be a member of the class of kinases, whose biological functions include phosphorylation of proteins. The claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires knowledge of how the polypeptide coded for by the polynucleotide actually functions.

Applicants submit with this brief the First Declaration of Bedilion describing some of the practical uses of the claimed invention in gene and protein expression monitoring applications. The First Bedilion Declaration demonstrates that the positions and arguments made by the Patent

Examiner with respect to the utility of the claimed polynucleotide are without merit.

The First Bedilion Declaration describes, in particular, how the claimed expressed polynucleotide can be used in gene expression monitoring applications that were well-known at the time the patent application was filed, and how those applications are useful in developing drugs and monitoring their activity. Dr. Bedilion states that the claimed invention is a useful tool when employed as a highly specific probe in a cDNA microarray:

> Persons skilled in the art would have appreciated on March 18, 1999 that cDNA microarrays that contained the SEQ ID NO:5-encoding polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating neurological, cell proliferative, and autoimmune/inflammatory disorders for such purposes as evaluating their efficacy and toxicity. (First Bedilion Declaration, ¶ 15.)

Applicants further submit three additional expert Declarations under 37 C.F.R. § 1.132, with respective attachments, and ten (10) scientific references filed before the March 18, 1999 priority date of the instant application. The First Bedilion Declaration, Rockett Declaration, Iyer Declaration, Second Bedilion Declaration, and the ten (10) references fully establish that, prior to the March 18, 1999 filing date of the parent application (Ser. No. 60/125,593, hereinafter the "Bandman '593 application"), it was well-established in the art that:

> polynucleotides derived from nucleic acids expressed in one or more tissues and/or cell types can be used as hybridization probes -- that is, as tools -- to survey for and to measure the presence, the absence, and the amount of expression of their cognate gene;
>
> with sufficient length, at sufficient hybridization stringency, and with sufficient wash stringency -- conditions that can be routinely established -- expressed polynucleotides, used as probes, generate a signal that is specific to the cognate gene, that is, produce a gene-specific expression signal;
>
> expression analysis is useful, *inter alia*, in drug discovery and lead optimization efforts, in toxicology, particularly toxicology studies conducted early in drug development efforts, and in phenotypic characterization and categorization of cell types, including neoplastic cell types;
>
> each additional gene-specific probe used as a tool in expression analysis provides an additional gene-specific signal that could not otherwise have been detected, giving a more comprehensive, robust, higher resolution,

statistically more significant, and thus more useful expression pattern in such analyses than would otherwise have been possible;

biologists, such as toxicologists, recognize the increased utility of more comprehensive, robust, higher resolution, statistically more significant results, and thus want each newly identified expressed gene to be included in such an analysis;

nucleic acid microarrays increase the parallelism of expression measurements, providing expression data analogous to that provided by older, lower throughput techniques, but at substantially increased throughput;

accordingly, when expression profiling is performed using microarrays, each additional gene-specific probe that is included as a signaling component on this analytical device increases the detection range, and thus versatility, of this research tool;

biologists, such as toxicologists, recognize the increased utility of such improved tools, and thus want a gene-specific probe to each newly identified expressed gene to be included in such an analytical device;

the industrial suppliers of microarrays recognize the increased utility of such improved tools to their customers, and thus strive to improve salability of their microarrays by adding each newly identified expressed gene to the microarrays they sell;

it is not necessary that the biological function of a gene be known for measurement of its expression to be useful in drug discovery and lead optimization analyses, toxicology, or molecular phenotyping experiments;

failure of a probe to detect changes in expression of its cognate gene does not diminish the usefulness of the probe as a research tool; and

failure of a probe completely to detect its cognate transcript in any single expression analysis experiment does not deprive the probe of usefulness to the community of users who would use it as a research tool.

The Patent Examiner does not dispute that the claimed polynucleotide can be used as a probe in cDNA microarrays and used in gene expression monitoring applications. Instead, the Patent Examiner contends that the claimed polynucleotide cannot be useful without precise knowledge of its biological function, or the biological function of the polypeptide it encodes. But the law has never required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the First Bedilion Declaration, the Rockett Declaration, the Iyer Declaration, and the Second Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotide in the absence of any knowledge as to the precise function of the protein encoded by it. The uses of the claimed polynucleotide in gene expression monitoring applications are in fact independent of its precise biological function.

I. The applicable legal standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

> An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end").

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. *See Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. *Cross* v. *Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." *Brenner*, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examining Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

II. The uses of polynucleotides encoding HRIP for diagnosis of conditions or diseases characterized by expression of HRIP and for drug discovery are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are "well-established" uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application's specification. These uses are explained, in detail, in the Bedilion Declaration accompanying this brief. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

A. The use of the claimed HRIP encoding polynucleotides for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer "specific benefits" to the public

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying First Bedilion Declaration, Rockett Declaration, Iyer Declaration, and Second Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotide.

The instant application is a U.S. National Stage of International Application No. PCT/US00/07277 and claims priority to a provisional application, Bandman et al., Ser. No. 60/125,593, filed on March 18, 1999, (hereinafter "the Bandman '593 application").

In his first Declaration, Dr. Bedilion explains the many reasons why a person skilled in the art reading the Bandman '593 application on March 18, 1999 would have understood that application to disclose the claimed polynucleotide to be useful for a number of gene expression monitoring applications, *e.g.*, as a highly specific probe for the expression of that specific polynucleotide in connection with the development of drugs and the monitoring of the activity of such drugs (Bedilion Declaration at, e.g., ¶¶ 10-15). Much, but not all, of Dr. Bedilion's explanation concerns the use of the claimed polynucleotide in cDNA microarrays of the type first developed at Stanford University for evaluating the efficacy and toxicity of drugs, as well as for other applications (First Bedilion Declaration at, e.g., ¶¶ 12 and 15).[1]

[1] Dr. Bedilion also explained, for example, why persons skilled in the art would also appreciate, based on the Bandman '593 specification, that the claimed polynucleotide would be useful in connection with developing new drugs using technology, such as Northern analysis, that predated by many years the development of the cDNA technology (First Bedilion Declaration, ¶

In connection with his explanations, Dr. Bedilion states that the "Bandman '593 application would have led a person skilled in the art in March 1999 who was using gene expression monitoring in connection with working on developing new drugs for the treatment of neurological, cell proliferative, and autoimmune/inflammatory disorders [a] to conclude that a cDNA microarray that contained the SEQ ID NO:5-encoding polynucleotides would be a highly useful tool, and [b] to request specifically that any cDNA microarray that was being used for such purposes contain the SEQ ID NO:5-encoding polynucleotides" (Bedilion Declaration, ¶ 15). For example, as explained by Dr. Bedilion, "[p]ersons skilled in the art would [have appreciated on March 18, 1999] that a cDNA microarray that contained the SEQ ID NO:5-encoding polynucleotides would be a more useful tool than a cDNA microarray that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating neurological, cell proliferative, and autoimmune/inflammatory disorders for such purposes as evaluating their efficacy and toxicity." *Id.*

In support of those statements, Dr. Bedilion provided detailed explanations of how cDNA technology can be used to conduct gene expression monitoring evaluations, with extensive citations to pre-March 18, 1999 publications showing the state of the art on March 18, 1999 (First Bedilion Declaration, ¶ ¶ 10-14). While Dr. Bedilion's explanations in paragraph 15 of his Declaration include almost three pages of text and six] subparts (a)-(f), he specifically states that his explanations are not "all-inclusive." *Id.* For example, with respect to toxicity evaluations, Dr. Bedilion had earlier explained how persons skilled in the art who were working on drug development on March 18, 1999 (and for several years prior to March 18, 1999) "without any doubt" appreciated that the toxicity (or lack of toxicity) of any proposed drug was "one of the most important criteria to be considered and evaluated in connection with the development of the drug" and how the teachings of the Bandman '593 application clearly include using differential gene expression analyses in toxicity studies (First Bedilion Declaration, ¶ 10).

Thus, the First Bedilion Declaration establishes that persons skilled in the art reading the Bandman '593 application at the time it was filed "would have wanted their cDNA microarray to have a [SEQ ID NO:5-encoding polynucleotide] probe because a microarray that contained

16).

such a probe (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies using cDNA microarrays that persons skilled in the art have been doing since well prior to March 18, 1999" (First Bedilion Declaration, ¶ 15, item (f)). This, by itself, provides more than sufficient reason to compel the conclusion that the Bandman '593 application disclosed to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the claimed polynucleotide.

In his Declaration, Dr. Rockett explains the many reasons why a person skilled in the art in 1997 would have understood that any expressed polynucleotide is useful for a number of gene expression monitoring applications, *e.g.*, in cDNA microarrays, in connection with the development of drugs and the monitoring of the activity of such drugs. (Rockett Declaration at, e.g., ¶¶ 10-18).

> It is my opinion, therefore, based on the state of the art in toxicology at least since the mid-1990s . . . that disclosure of the sequence of a new gene or protein, with or without knowledge of its biological function, would have been sufficient information for a toxicologist to use the gene and/or protein in expression profiling studies in toxicology.[2] [Rockett Declaration, ¶ 18.]

In his second Declaration, Dr. Bedilion explains why a person of skill in the art in 1997 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays. (Second Bedilion Declaration, e.g., ¶¶ 4-7.) In his Declaration, Dr. Iyer explains why a person of skill in the art in 1997 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays, stating that "[t]o provide maximum versatility as a research tool, the microarray should include □ and as a biologist I would want my microarray to include □ each newly identified gene as a probe." (Iyer Declaration, ¶ 9.)

In addition, Dr. Rockett explains in his Declaration that "there are a number of other differential expression analysis technologies that precede the development of microarrays, some by decades, and that have been applied to drug metabolism and toxicology research, including:

"Use of the words 'it is my opinion' to preface what someone of ordinary skill in the art would have known does not transform the factual statements contained in the declaration into opinion testimony." *In re Alton*, 37 USPQ2d 1578, 1583 (Fed. Cir. 1996).

(1) differential screening; (2) subtractive hybridization, including variants such as chemical cross-linking subtraction, suppression-PCR subtractive hybridization and representational difference analysis; (3) differential display; (4) restriction endonuclease facilitated analyses, including serial analysis of gene expression (SAGE) and gene expression fingerprinting and (5) EST analysis." (Rockett Declaration, ¶ 7.)

Nowhere does the Patent Examiner address the fact that, as described on, for example, page 34 of the Bandman '593 application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays □ probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon* v. *Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 ("Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)" (emphasis added)).

The First Bedilion Declaration shows that a number of pre-March 18, 1999 publications confirm and further establish the utility of cDNA microarrays in a wide range of drug development gene expression monitoring applications at the time the Bandman '593 application was filed (First Bedilion Declaration ¶¶ 10-14; Bedilion Exhibits A-G). Indeed, Brown and Shalon U.S. Patent No. 5,807,522 (the Brown '522 patent, Bedilion Exhibit D), which issued from a patent application filed in June 1995 and was effectively published on December 29, 1995 as a result of the publication of a PCT counterpart application, shows that the Patent Office recognizes the patentable utility of the cDNA technology developed in the early to mid-1990s. As explained by Dr. Bedilion, among other things (First Bedilion Declaration, ¶ 12):

> The Brown '522 patent further teaches that the "[m]icroarrays of immobilized nucleic acid sequences prepared in accordance with the invention" can be used in "numerous" genetic applications, including "monitoring of gene expression" applications (see Bedilion Tab D at col. 14, lines 36-42). The Brown '522 patent teaches (a) monitoring gene expression (i) in different tissue types, (ii) in different disease states, and (iii) in response to different drugs, and (b) that arrays disclosed therein may be used in toxicology studies (see Bedilion Tab D at col. 15, lines 13-18 and 52-58; and col. 18, lines 25-30).

Literature reviews published shortly after the filing of the Bandman '593 application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

> Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.
>
> * * *
>
> Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.
>
> * * *
>
> Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original)

Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, Xenobiotica 29:655-691 (July 1999) (Reference No. 2).

In another pre-March 18, 1999 article, Lashkari et al. state explicitly that sequences that are merely "predicted" to be expressed (predicted Open Reading Frames, or ORFs) □ the claimed invention in fact is known to be expressed □ have numerous uses:

> Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons□ they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay. (emphasis added)

Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (Aug. 1997) (Reference No. 3).

B. The use of polynucleotides coding for polypeptides expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now "well-established"

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, e.g., as described by Bedilion, Rockett, and Iyer in their Declarations.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., *supra*:

> Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Reference No. 2, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999) (Reference No. 4); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000) (Reference No. 5).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

> for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, Environ. Health Perspec.107:681-685 (1999) (Reference No. 6). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 7), indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

Further evidence of the well-established utility of all expressed polypeptides and polynucleotides in toxicology testing is found in U.S. Pat. No. 5,569,588 (Reference No. c) and published PCT applications WO 95/21944 (Reference No. a), WO 95/20681 (Reference No. b), and WO 97/13877 (Reference No. d).

WO 95/21944 ("Differentially expressed genes in healthy and diseased subjects"), published August 17, 1995, describes the use of microarrays in expression profiling analyses, emphasizing that *patterns* of expression can be used to distinguish healthy tissues from diseased tissues and that *patterns* of expression can additionally be used in drug development and toxicology studies, without knowledge of the biological function of the encoded gene product. In particular, and with emphasis added:

> The present invention involves . . . methods for diagnosing diseases . . . characterized by the presence of [differentially expressed] . . . genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/ polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST. . . . Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of diseases based on differential expression of genes of unknown function . . . . [abstract]
>
> The method [of the present invention] involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences . . . and a defined set of oligonucleotide/polynucleotide[] . . . immobilized on a support. Those defined [immobilized] oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined by the collection of partial cDNA sequences (ESTs). [page 2]
>
> The present invention meets the unfilled needs in the art by providing methods for the . . . use of gene fragments and genes, even those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. [page 4]
>
> Yet another aspect of the invention is that it provides . . . a means for . . . monitoring the efficacy of disease treatment regimes including . . . toxicological effects thereof." [page 4]
>
> It has been appreciated that one or more differentially identified EST or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the EST[] identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. [page 4]
>
> As used herein, the term 'disease' or 'disease state' refers to any condition which deviates from a normal or standardized healthy state in an organism of the

same species in terms of differential expression of the organism's genes. . . [whether] of genetic or environmental origin, for example, an inherited disorder such as certain breast cancers. . . .[or] administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects gene expression. [page 5]

As used herein, the term 'solid support' refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method . . . [and includes, inter alia,] nitrocellulose, . . . glass, silica. . . . [page 6]

By 'EST' or 'Expressed Sequence Tag' is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides. . . . [page 6]

One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000 100,000 in an animal such as a human. [page 7]

The lengths of the defined oligonucleotide/ polynucleotides may be readily increased or decreased as desired or needed. . . . The length is generally guided by the principle that it should be of sufficient length to insure that it is on[] average only represented once in the population to be examined. [page 7]

Comparing the . . . hybridization patterns permits detection of those defined oligonucleotide/ polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions [of the solid support]. [page 13]

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained[;] rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used [to] form diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a 'master' solid support to which control sample and disease samples are alternately hybridized. [page 14]

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This[,] like many of the foregoing embodiments[,] may use known or unknown ESTs derived from many libraries. [page 18]

> Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal[s], especially humans. [page 18]

WO 95/20681 ("Comparative Gene Transcript Analysis"), filed in 1994 by Applicants' assignee and published August 3, 1995, has three issued U.S. counterparts: U.S. Pat. Nos. 5,840,484, issued November 24, 1998; 6,114,114, issued September 5, 2000; and 6,303,297, issued October 16, 2001.

The specification describes the use of transcript expression *patterns,* or "images", each comprising multiple pixels of gene-specific information, for diagnosis, for cellular phenotyping, and in toxicology and drug development efforts. The specification describes a plurality of methods for obtaining the requisite expression data -- one of which is microarray hybridization -- and equates the uses of the expression data from these disparate platforms. In particular, and with emphasis added:

> The invention provides a "method and system for quantifying the relative abundance of gene transcripts in a biological specimen. . . . [G]ene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens." [abstract]
>
> "[W]e see each individual gene product as a 'pixel' of information which relates to the expression of that, and only that, gene. We teach herein [] methods whereby the individual 'pixels' of gene expression information can be combined into a single gene transcript 'image,'in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood." [page 2]
>
> "The present invention avoids the drawbacks of the prior art by providing a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen. . . . The method of the instant invention provides for detailed diagnostic comparisons of cell profiles revealing numerous changes in the expression of individual transcripts." [page 6]
>
> "High resolution analysis of gene expression be used directly as a diagnostic profile. . . . " [page 7]

"The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed." [page 7]

"The invention . . . includes a method of comparing specimens containing gene transcripts." [page 7]

"The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens." [i.e., the results yield analogous data to microarrays] [page 8]

"Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made." [page 8]

"In a further embodiment, the relative abundance o the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities." [page 9]

"In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens. . . . " [page 9]

"[T]wo or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells." [pages 9 10]

"The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens. . . . This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as 'gene transcript image analysis' or 'gene transcript frequency analysis'. The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism." [page 11]

"The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few." [page 12]

"[G]ene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patent has the disease(s) with which the patient'sdata set most closely correlates." [page 12]

"For example, gene transcript frequency analysis can be used to different normal cells or tissues from diseased cells or tissues. . . ." [page 12]

"In toxicology, . . . [g]ene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. . . . " [page 12]

"In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond." [page 12]

"In a further embodiment, comparative gene transcript frequency analysis is used . . . for the selection of better pharmacologic animal models." [page 14]

"In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a diseased state or condition." [page 14]

"An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined." [page 15]

"[T]his research tool provides a way to get new drugs to the public faster and more economically." [page 36]

"In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker." [page 38]

"[T]he gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript image analyses are evaluated as indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. . . . The . . . analysis highlights any toxicological changes in the treated patients." [page 39]

U.S. Pat. No. 5,569,588 ("Methods for Drug Screening") ("the '588 patent"), issued October 29, 1996, with a priority date of August 1995, describes an expression profiling platform, the "genome reporter matrix", which is different from nucleic acid microarrays. Additionally describing use of nucleic acid microarrays, the patent makes clear that the utility of comparing multidimensional expression datasets is independent of the methods by which such profiles are obtained. The patent speaks clearly to the usefulness of such expression analyses in drug development and toxicology, particularly pointing out that a gene's failure to change in expression level is a useful result. Thus, with emphasis added,

> The invention provides "[m]ethods and compositions for modeling the transcriptional responsiveness of an organism to a candidate drug. . . . [The final step of the method comprises] comparing reporter gene product signals for each cell before and after contacting the cell with the candidate drug to obtain a drug response profile which provides a model of the transcriptional responsiveness of said organism to the candidate drug." [abstract]
>
> "The present invention exploits the recent advances in genome science to provide for the rapid screening of large numbers of compounds against a systemic target comprising substantially all targets in a pathway [or] organism." [col. 1]
>
> "The ensemble of reporting cells comprises as comprehensive a collection of transcription regulatory genetic elements as is conveniently available for the targeted organism so as to most accurately model the systemic transcriptional response. Suitable ensembles generally comprise thousands of individually reporting elements; preferred ensembles are substantially comprehensive, i.e. provide a transcriptional response diversity comparable to that of the target organism. Generally, a substantially comprehensive ensemble requires transcription regulatory genetic elements from at least a majority of the organism's genes, and preferably includes those of all or nearly all of the genes. We term such a substantially comprehensive ensemble a genome reporter matrix." [col. 2]
>
> "Drugs often have side effects that are in part due to the lack of target specificity. . . . [A] genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reports, the first compound is, a priori, more likely to have fewer side effects." [cols. 2 - 3]
>
> "Furthermore, it is not necessary to know the identity of any of the responding genes." [col. 3]

"[A]ny new compound that induces the same response profile as [a] . . . dominant tubulin mutant would provide a candidate for a taxol-like pharmaceutical." [col. 4]

"The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters." [col. 4]

" A sufficient number of different recombinant cells are included to provide an ensemble of transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug. In a preferred embodiment, the matrix is substantially comprehensive for the selected regulatory elements, e.g. essentially all of the gene promoters of the targeted organism are included." [cols. 6 7]

"In a preferred embodiment, the basal response profiles are determined. . . . The resultant electrical output signals are stored in a computer memory as genome reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus or drug. This information is indexed against the matrix to form reference response profiles that are used to determine the response of each reporter to any milieu in which a stimulus may be provided. After establishing a basal response profile for the matrix, each cell is contacted with a candidate drug. The term drug is used loosely to refer to agents which can provoke a specific cellular response. . . . The drug induces a complex response pattern of repression, silence and induction across the matrix . . . .The response profile reflects the cell's transcriptional adjustments to maintain homeostasis in the presence of the drug. . . . After contacting the cells with the candidate drug, the reporter gene product signals from each of said cells is again measured to determine a stimulated response profile. The basal o[r] background response profile is then compared with . . . the stimulated response profile to identify the cellular response profile to the candidate drug." [cols. 7 8]

"In another embodiment of the invention, a matrix [i.e., array] of hybridization probes corresponding to a predetermined population of genes of the selected organism is used to specifically detect changes in gene transcription which result from exposing the selected organism or cells thereof to a candidate drug. In this embodiment, one or more cells derived from the organism is exposed to the candidate drug in vivo or ex vivo under conditions wherein the drug effects a change in gene transcription in the cell to maintain homeostasis. Thereafter, the gene transcripts, primarily mRNA, of the cell or cells is isolated . . . [and] then contacted with an ordered matrix [array] of hybridization probes, each probe being specific for a different one of the transcripts, under conditions where each of the transcripts hybridizes with a corresponding one of the probes to form hybridization pairs. The ordered matrix of probes provides, in aggregate, complements for an

> ensemble of genes of the organism sufficient to model the transcriptional responsiveness of the organism to a drug. . . . The matrix-wide signal profile of the drug-stimulated cells is then compared with a matrix-wide signal profile of negative control cells to obtain a specific drug response profile." [col. 8]
>
> "The invention also provides means for computer-based qualitative analysis of candidate drugs and unknown compounds. A wide variety of reference response profiles may be generated and used in such analyses." [col. 8]
>
> "Response profiles for an unknown stimulus (e.g. new chemicals, unknown compounds or unknown mixtures) may be analyzed by comparing the new stimulus response profiles with response profiles to known chemical stimuli." [col. 9]
>
> "The response profile of a new chemical stimulus may also be compared to a known genetic response profile for target gene(s)." [col. 9]

The August 11, 1997 press release from the '588 patent's assignee, Acacia Biosciences (now part of Merck) (reference "h" attached hereto), and the September 15, 1997 news report by Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," *Genetic Engineering News* (reference "i" attached hereto), attest the commercial value of the methods and technology described and claimed in the '588 patent.

WO 97/13877 ("Measurement of Gene Expression Profiles in Toxicity Determinations"), published April 17, 1997; describes an expression profiling technology differing somewhat from the use of cDNA microarrays and differing from the genome reporter matrix of the '588 patent; but the use of the data is analogous. As per its title, the reference describes use of expression profiling in toxicity determinations. In particular, and with emphasis added:

> "[T]he invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates." [Field of the invention]
>
> "An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems." [page 3]
>
> "Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals." [page 3]

> "The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel . . . methodologies that permit the formation of gene expression profiles for selected tissues . . . . Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity." [page 3]
>
> "As used herein, the terms 'gene expression profile,' and 'gene expression pattern' which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. . .. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences determined in a gene expression profile is at least ten thousand." [page 7]
>
> "The invention provides a method for determining the toxicity of a compound by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. . . . . Gene expression profiles derived from test organisms are compared to gene expression profiles derived from control organisms. . . . " [page 7]

Therefore, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Evidence of the benefits of this information include:

□ In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

□ In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.

□ In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be overturned regardless of their merit.

C. The similarity of the polypeptide encoded by the claimed invention to another polypeptide of undisputed utility demonstrates utility

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, the utility of the claimed polynucleotide can be imputed based on the relationship between the polypeptide it encodes, HRIP, and another polypeptide of unquestioned utility, sphingosine kinase. The two polypeptides have sufficient similarities in their sequences that a person of ordinary skill in the art would recognize more than a reasonable probability that the polypeptide encoded for by the claimed invention has utility similar to sphingosine kinase. Applicants need not show any more to demonstrate utility. *In re Brana*, 51 F.3d at 1567.

It is undisputed, and readily apparent from the patent application, that the polypeptide encoded for by the claimed polynucleotide shares more than 80% sequence identity over 384 amino acid residues with mouse sphingosine kinase (g3659694). Furthermore, an alignment of SEQ ID NO:5 with a post-filing human sphingosine kinase shows that the two sequences are approximately 99% identical over the entire 384 amino acid residue length of both sequences. In addition, a diacylglycerol kinase catalytic domain was identified by searching for statistically significant matches in the hidden Markov model (HMM)-based PFAM database of conserved protein families/domains. Sequence analysis of several families of kinases suggests that diacylglycerol kinases and sphingosine kinases are members of the same superfamily of genes due to a common domain (Labesse *et al.*, Trends Biochem. Sci. 27:273-5, 2002; Reference No. 8). In an earlier report, Kohama *et al.* state that "the C1 and C3 subdomains of sphingosine kinase show high amino acid similarity to residues 296-315 and 378-389 of human diacylglycerol kinase ζ with 35% and 58% identity, respectively" (JBC 273:23722-8, 1998). Taken together, these data suggest that there is more than enough homology to demonstrate a reasonable probability that the utility of sphingosine kinase can be imputed to the claimed invention (through the polypeptide it encodes). It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. (Brenner et al., Proc. Natl. Acad. Sci. 95:6073-78 (1998); Reference No. 9.) Given homology in excess of 40% over many more than 70 amino acid residues, the probability that the polypeptide encoded for by the claimed polynucleotide is related to mouse sphingosine kinase is, accordingly, very high.

The Examiner must accept the Applicants' demonstration that the homology between the polypeptide encoded for by the claimed invention and mouse sphingosine kinase demonstrates utility by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. See *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

D. Objective evidence corroborates the utilities of the claimed invention

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp.* v. *Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a thriving market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Applicants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's databases have proven to be valuable in, for example, the identification and development of drug candidates. Page et al., in discussing the identification and assignment of candidate drug targets, state that "rapid identification and assignment of candidate targets and markers represents a huge challenge ... [t]he process of annotation is similarly aided by the quantity and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals)" Page, M.J. et al., "Proteomics: a major new technology for the drug discovery process," Drug Discov. Today 4:55-62 (1999) (Reference No. 10), see page 58, col. 2). As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's invention of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

Customers can, moreover, purchase the claimed polynucleotide directly from Incyte, saving the customer the time and expense of isolating and purifying or cloning the polynucleotide for research uses such as those described *supra.*

III. The Patent Examiner's rejections are without merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that "disclosure that a protein is a kinase without a more specific recitation of what type of kinase (i.e., what compound(s) is phosphorylated)", the disclosed and well-established utilities for the claimed polynucleotide are not specific or substantial utilities (Office Action at page 7). The Examiner is incorrect both as a matter of law and as a matter of fact.

A. The precise biological role or function of an expressed polynucleotide is not required to demonstrate utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc.* v. *Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Bedilion Declaration (at, *e.g.*, ¶¶ 10 and 15), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

> [T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

B. Membership in a class of useful products can be proof of utility

Despite the evidence that the claimed polynucleotide encodes a polypeptide in the kinase family, the Examiner refused to impute the utility of the members of the kinase family to HRIP. In the Office Action, the Patent Examiner takes the position that, unless Applicants can identify which particular biological function within the class of kinases is possessed by HRIP (i.e., the substrate of HRIP), utility cannot be imputed. To demonstrate utility by membership in the class of kinases, the Examiner would require that all kinases possess a "common" utility. The Examiner is incorrect both as a matter of fact and of law.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. See *Brenner* v. *Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a "general" class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses HRIP as if the general class in which it is included is not the kinase family, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these "general classes" may contain a substantial number of useless members, the kinase family does not. The kinase family is sufficiently specific to rule out any reasonable possibility that HRIP would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the kinase family has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a "substantial likelihood" that the HRIP encoded by the claimed polynucleotide is useful. It follows that the claimed polynucleotide also is useful.

Even if the Examiner's "common utility" criterion were correct - and it is not - the kinase family would meet it. It is undisputed that known members of the kinase family phosphorylate proteins. A person of ordinary skill in the art need not know any more about how or what the claimed invention phosphorylates to use it, and the Examiner presents no evidence to the contrary. Instead, the Examiner makes the conclusory observation that a person of ordinary skill in the art would need to know the substrate of any given kinase.

Not so. As demonstrated by Applicants, knowledge that HRIP is a kinase is more than sufficient to make it useful for the diagnosis and treatment of neurological, cell proliferative, and autoimmune/inflammatory disorders. Indeed, HRIP has been shown to be expressed in cells undergoing proliferation or involved in inflammation (see the specification at, for example, p. 62). The Examiner must accept these facts to be true unless the Examiner can provide evidence or sound scientific reasoning to the contrary. But the Examiner has not done so.

IV. By requiring the patent applicant to assert a particular or unique utility, the Patent Examination Utility Guidelines and Training Materials applied by the Patent Examiner misstate the law

There is an additional, independent reason to overturn the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities which meet the statutory requirements, and "general" utilities which do not.

The Training Materials define a "specific utility" as follows:

> A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between "specific" and "general" utilities by assessing whether the asserted utility is sufficiently "particular," *i.e.*, unique (Training Materials at page 52) as compared to the "broad class of invention." (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) ("With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.")).

Such "unique" or "particular" utilities never have been required by the law. To meet the utility requirement, the invention need only be "practically useful," *Natta*, 480 F.2d 1 at 1397, and confer a "specific benefit" on the public. *Brenner*, 383 U.S. at 534. Thus, incredible "throw-away" utilities, such as trying to "patent a transgenic mouse by saying it makes great snake food," do not meet this standard. Karen Hall, Genomic Warfare, The American Lawyer 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where "specific utility" is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be "definite," not particular. *Montedison,* 664 F.2d at 375. Applicants are not aware of any court that has rejected an assertion of utility on the grounds that it is not "particular" or "unique" to the specific invention. Where courts have found utility to be too "general," it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had "useful biological activity" was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376

F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a "particular" type of cancer was determined to satisfy the specificity requirement). "Particularity" is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § III.B.2 (*Montedison,* 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of "general" utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. See *supra* § II.B. Thus the Training Materials cannot be applied consistently with the law.

V. To the extent the rejection of the claimed invention under 35 U.S.C. § 112, first paragraph, is based on the improper rejection for lack of utility under 35 U.S.C. § 101, it must be reversed.

The rejection set forth in the Office Action is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under 35 U.S.C. § 112, first paragraph, is based on the improper allegation of lack of patentable utility under 35 U.S.C. § 101, it fails for the same reasons.

VI. Summary

Applicants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of "lack of specificity," as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, "like a nose of wax,"[3] to target rejections of claims to polypeptide and polynucleotide sequences, as well as to claims to methods of detecting said polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be reversed.

Enablement rejection under 35 U.S.C. §112, 1st paragraph

Claims 3, 5-6, 8, and 10-11 are rejected under 35 U.S.C. §112, first paragraph, as allegedly containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to make and/or use the invention. The Examiner asserts that the specification does not support the breadth of the claims with respect to naturally occurring variants, biologically active fragments, immunologically active fragments, and polynucleotide fragments of at least 60 nucleotides. (Applicants note that claim 11 has been amended and now recites fragments of at least 500 nucleotides.) Applicants traverse this rejection for at least the reasons below.

As set forth in *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971):

> The first paragraph of § 112 requires nothing more than **objective enablement**. How such a teaching is set forth, either by the use of illustrative examples or by broad terminology, is of no importance.

[3] "The concept of patentable subject matter under §101 is not 'like a nose of wax which may be turned and twisted in any direction * * *.' *White v. Dunbar*, 119 U.S. 47, 51." *(Parker v. Flook,* 198 USPQ 193 (US SupCt 1978))

> As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.

Applicants submit that the disclosure amply enables the claimed invention. Given the sequences of SEQ ID NO:5 and SEQ ID NO:19, one of ordinary skill in the art could readily identify a polynucleotide encoding a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:5 or a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical to a polynucleotide sequence of SEQ ID NO:19, using well known methods of sequence analysis without any undue experimentation. For example, the identification of relevant polynucleotides could be performed by hybridization and/or PCR techniques that were well-known to those skilled in the art at the time the subject application was filed and/or described throughout the Specification of the instant application. See, e.g., p. 42, lines 17-24; and Example VI at p. 52, lines 5-21. Thus, one skilled in the art need not make and test vast numbers of polynucleotides. Instead, one skilled in the art need only screen a cDNA library or use appropriate PCR conditions to identify relevant polynucleotides that already exist in nature. The skilled artisan would also know how to use the claimed polynucleotides, for example in expression profiling, disease diagnosis, or detection of related sequences as discussed above.

The specification also describes the expression vectors into which the claimed variants and fragments could be inserted, and the construction of fusion proteins (pages 31-32). The specification describes, binding assays to detect molecular interactions of "HRIP or biologically active fragments thereof" on page 56; and immunological methods for detecting and measuring HRIP on, for example, page 56. These methods could be used to detect and characterize peptide variants and fragments of SEQ ID NO:5. Given this guidance, one of ordinary skill in the art would readily understand how to select and screen polynucleotides encoding variants or fragments of SEQ ID NO:5 without any undue experimentation.

To expedite prosecution, claim 3 has been amended to recite "a biologically active fragment comprising at least 150 contiguous amino acids of the amino acid sequence of SEQ ID NO:5, wherein said biologically active fragment has sphingosine kinase activity." Applicants are amending the claim solely to obtain expeditious allowance of the instant application. Support for this amendment to claim 3 can be found in the specification, for example, in Table 2 which points out the homology between SEQ ID NO:5 and mouse sphingosine kinase (g3659694), and at p. 54, lines 11-24, which describes assays for measuring kinase activity. By this amendment, Applicants expressly do not disclaim equivalents of the invention which could include polypeptides or fragments having biological activities in addition to sphingosine kinase inducing activity.

The Examiner suggests that, in order to satisfy the enablement requirement, knowledge of the regions of SEQ ID NO:5 that are tolerant to modification, the tolerance of kinases in general to modification, a scheme for modifying SEQ ID NO:5 while obtaining the desired biological function, and guidance as to which were likely to successful is required. Applicants wish to point out that each of these criteria are directed to the function of the polypeptide not the polynucleotide. Applicants respectfully remind the Examiner that the claims are directed to polynucleotides, not polypeptides, and thus it is the functionality of the claimed polynucleotides, not the polypeptides encoded by them, that is relevant.

With respect to the claimed variants of SEQ ID NO:19, members of this genus may, for example, be useful even if they encode proteins that lack sphingosine kinase activity. For example, the variant polynucleotides could be used for the detection of sequences related to sphingosine kinase (see the specification at p. 42, lines 25-28) including sphingosine kinase variants that may be associated with disease states, such as the diseases listed in the specification at p. 43, line 2 through p. 44, line 5). See the specification at, for example, p. 44, lines 19-27 for disclosure of how to use the claimed sequences in diagnostic assays. The variant polynucleotides could also be used in microarrays to identify genetic variants, mutations, and polymorphisms, and for disease diagnosis and development and testing of therapeutic agents (see the specification at, for example, p. 45, lines 18-28). Thus one of ordinary skill in the art would know how to used the claimed polynucleotide variants without having to experimentally determine the biological function of the encoded proteins because the function or lack thereof of the protein is irrelevant.

Contrary to the Examiner's assertions, immunogenic fragments of SEQ ID NO:5 are amply enabled by the disclosure of the specification. For example, at page, lines, the specification describes methods for identifying immunogenic fragments.

> "Alternatively, the HRIP amino acid sequence is analyzed using LASERGENE software (DNASTAR) to determine regions of high immunogenicity, and a corresponding oligopeptide is synthesized and used to raise antibodies by means known to those of skill in the art. Methods for selection of appropriate epitopes, such as those near the C-terminus or in hydrophilic regions are well described in the art. (See, e.g., Ausubel, 1995, supra, ch. 11.)"
> (Specification at page 55, lines 24-28)

The specification further describes the use of immunogenic fragments to induce antibodies that bind specifically to a given region of a protein.

> "When a protein or a fragment of a protein is used to immunize a host animal, numerous regions of the protein may induce the production of antibodies which bind specifically to antigenic determinants (particular regions or three-dimensional structures on the protein). An antigenic determinant may compete with the intact antigen (i.e., the immunogen used to elicit the immune response) for binding to an antibody."
> (Specification at page 11, lines 28-32)

At page 35, lines 5-11, the specification states:

> "For the production of antibodies, various hosts including goats, rabbits, rats, mice, humans, and others may be immunized by injection with HRIP or with any fragment or oligopeptide thereof which has immunogenic properties. Depending on the host species, various adjuvants may be used to increase immunological response. Such adjuvants include, but are not limited to, Freund's, mineral gels such as aluminum hydroxide, and surface active substances such as lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, KLH, and dinitrophenol. Among adjuvants used in humans, BCG (bacilli Calmette-Guerin) and Corynebacterium parvum are especially preferable."

The specification continues at page 55, lines 29-35 with a description of the immunogenic fragments that could be used to induce antibodies:

> "Typically, oligopeptides of about 15 residues in length are synthesized using an ABI 431A peptide synthesizer (Perkin-Elmer) using fmoc-chemistry and coupled to KLH (Sigma-Aldrich, St. Louis MO) by reaction with N-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS) to increase immunogenicity. (See, e.g., Ausubel, 1995, supra.) Rabbits are immunized with the oligopeptide-KLH complex in complete Freund's

> adjuvant. Resulting antisera are tested for antipeptide and anti-HRIP activity by, for example, binding the peptide or HRIP to a substrate, blocking with 1% BSA, reacting with rabbit antisera, washing, and reacting with radio-iodinated goat anti-rabbit IgG."

Immunogenic fragments of SEQ ID NO:5 by definition elicit antibodies that bind to SEQ ID NO:5. It is routine to produce antibodies that specifically bind to a protein by immunizing an appropriate host with oligopeptide fragments of a protein. It is well known in the art that it is possible to produce antibodies to almost any part of an antigen. Given the sequence of SEQ ID NO:5, one of skill in the art could readily identify immunogenic fragments of SEQ ID NO:5.

Contrary to the standard set forth in *Marzocchi* and *Borkowski*, the Examiner has failed to provide any ***reasons*** why one would doubt that the guidance provided by the present specification would enable one to make and use the recited polynucleotides. Hence, a *prima facie* case for non-enablement has not been established. For at least the above reasons, withdrawal of the enablement rejection under 35 U.S.C. § 112, first paragraph, is respectfully requested.

Written description rejection under 35 U.S.C. §112, 1st paragraph

Claims 3, 5-6, 8, and 10-11 have been rejected under the first paragraph of 35 U.S.C. 112 for alleged lack of an adequate written description. This rejection is respectfully traversed. The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

> . . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention.* The invention is, for purposes of the "written description" inquiry, *whatever is now claimed. Vas-Cath, Inc. v. Mahurkar,* 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that :

> An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics

> when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met. (footnotes omitted.)

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:5 and SEQ ID NO:19 (the polynucleotide sequence encoding SEQ ID NO:5) are specifically disclosed in the application (see, for example, page 6, lines 14-30). Variants of SEQ ID NO:19 are described, for example, at page 24, lines 8-16. In particular, the variants of SEQ ID NO:19 are described in the alternative as at least 80%, at least 90%, or at least 95% sequence identity to a polynucleotide sequence having SEQ ID NO:19 at, for example, page 24, lines 11-16. Incyte clones in which the nucleic acids encoding the human sphingosine kinase were first identified and libraries from which those clones were isolated are described, for example, at pages 64-65 of the specification. Chemical and structural features of the protein encoded by SEQ ID NO:19 are described, for example, on page 59, row 6. Given SEQ ID NO:19, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:19 having 90% sequence identity to SEQ ID NO:19. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

The Office Action has further asserted that the claims are not supported by an adequate written description because "many functionally unrelated DNAs are encompassed within the scope of these claims" (page 12 of the Office Action of June 4, 2003). Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35

U.S.C. §112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides in terms of chemical structure, rather than functional characteristics. For example, the "variant language" of independent claim 10, as presently amended, recites chemical structure to define the claimed genus:

> 10. An isolated polynucleotide comprising a polynucleotide sequence selected from the group consisting of:...b) a naturally occurring polynucleotide sequence having at least 90% sequence identity to a polynucleotide sequence of SEQ ID NO:19,

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:19. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides. The polynucleotides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*

2. The present claims do not define a genus which is "highly variant"

Furthermore, the claims at issue do not describe a genus which could be characterized as "a large variable genus." Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the enclosed reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia,* to kinase proteins related to the amino acid sequence of SEQ ID NO:5. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as kinase proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:5. The "variant language" of the present claims recites, for example, polynucleotides encoding "a naturally-occurring amino acid sequence having at least 90% sequence identity to the amino acid sequence of SEQ ID NO:5" (note that SEQ ID NO:5 has 384 amino acid residues). This variation is far less than that of all potential kinase proteins related to SEQ ID NO:5, i.e., those kinase proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:.

3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of March 18, 1999. Much has happened in the development of recombinant DNA technology in the 24 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:5 and SEQ ID NO:19, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

4. Summary

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:5 or SEQ ID NO:19. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides defined by the present claims is adequately described, as evidenced by Brenner et al and consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action. Applicants respectfully request that this rejection be withdrawn.

Rejections under 35 U.S.C. §102(b)

Claims 3 and 11 are rejected under 35 U.S.C. §102(b) as allegedly being anticipated by Genbank accession # AA639414. These claims as presently amended, recite fragments comprising at least 150 contiguous amino acid residues or at least 500 contiguous nucleotides. This rejection has therefore been rendered moot. Withdrawal of this rejection is respectfully requested.

Claims 3, 5-6, 8, and 10-11 are rejected under 35 U.S.C. §102(b) as allegedly being anticipated by Kohama *et al.* Applicants respectfully point out that Kohama *et al.* (JBC 273:23722-8, 1998) was published on September 11, 1998, less than one year prior to Applicants' priority date of March 18, 1999. Therefore this reference does not qualify as prior art under 35 U.S.C. §102(b). Applicants respectfully request that this rejection be withdrawn.

Rejections under 35 U.S.C. §102(a)

Claims 3 and 11 are rejected under 35 U.S.C. § 102(a) as allegedly being anticipated by Genbank accession # AI042283, published September 24, 1998. This reference, however, was published after Applicants' date of invention. Attached is Reference No. 11, which demonstrates the date of invention of SEQ ID NO:5 as at least as early as December 13, 1996. Reference 11 is a print-out of the clone information for SEQ ID NO:19 (Project ID 2415617), showing an "initial entry date" of December 13, 1996. This antedates this Genbank reference, thus removing it as prior art. Therefore, the Examiner has failed to make out a prima facie case because she has not cited a single prior art reference that discloses all of the elements and limitations of Applicants' present claims. Applicants respectfully request that the rejection be withdrawn.

Claims 3, 5-6, and 8 are rejected under 35 U.S.C. § 102(a) as allegedly being anticipated by Young *et al.* (WO98/54963). The priority filing date of this reference is June 6, 1997 which is after Applicants' date of invention of SEQ ID NO:5 of December 13, 1996, as established above. Applicants' date of invention antedates the Young *et al.* reference, thus eliminating it as prior art. Therefore, the Examiner has again failed to make out a prima facie case because she has not cited a single prior art reference that discloses all of the elements and limitations of Applicants' present claims. Applicants respectfully request that the rejection be withdrawn.

Rejection under 35 U.S.C. §103(a)

Claim 10 is rejected under 35 U.S.C. §103(a) as allegedly being obvious over Kohama *et al.* in view of Genbank accession numbers D31133, AA232791, W63556, AA081152, and AA026479. However, the Examiner has not met her burden of making a *prima facie* case of obviousness for at least the reasons given below.

As stated previously, the Kohama *et al.* reference does not qualify as prior art under 35 U.S.C. §102(b) because its date of publication is less than one year prior to the filing date of the present application. Furthermore, this reference does not qualify as prior art under 35 U.S.C. §102(a) because Applicants' date of invention of December 13, 1996 antedates the Kohama *et al.* reference. Moreover, even assuming that all of the EST's combined in the Kohama *et al.* reference to arrive at the putative human sequence contained therein could be asserted as prior art, there is no motivation to combine absent knowledge of Applicants' invention or the mouse sphingosine kinase sequence. This is impermissible hindsight reconstruction. Without the motivation to combine the Examiner has not met the burden of establishing a *prima facie* case of obviousness. Withdrawal of this rejection is therefore respectfully requested.

CONCLUSION

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding objections/rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact the undersigned at the number listed below.

Please charge Deposit Account No. **09-0108** in the amount of **$<u>1130.00</u>** as set forth in the enclosed fee transmittal letter. If the USPTO determines that an additional fee is necessary, please charge any required fee to Deposit Account No. 09-0108.

Respectfully submitted,
INCYTE CORPORATION

Date: Dec 4, 2003

[illegible], Reg # 41,113
for: Cathleen M. Rocco
Reg. No. 46,172
Direct Dial Telephone: (650) 845-4587

Date: 04 December 2003

Karin M. Gerstin
Reg. No. 54,119
Direct Dial Telephone: (650) 845-4889

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

<u>References Enclosed</u>:

1. Nava et al., <u>Functional characterization of human sphingosine kinase-1</u>, FEBS Letters 473:81-4 (2000).
2. Rockett et al., <u>Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential</u>, Xenobiotica 29:655-691 (1999).

3. Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (1997).
4. Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999);
5. Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).
6. John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, 107 Environ. Health Perspec. 107:681-685 (1999).
7. Email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.
8. Labesse et al., Diacylglyceride kinases, sphingosine kinases and NAD kinases: distant relatives of 6-phosphofructokinases, Trends in Biochem. Sci. 27:273-5 (2002).
9. Brenner et al., Proc. Natl. Acad. Sci. 95:6073-6078 (1998).
10. Page, M.J. et al., Proteomics: a major new technology for the drug discovery process, Drug Discov. Today 4:55-62 (1999).
11. Print-out of Clone Information for SEQ ID NO:19 (Project ID 2415617).

Declarations and References Enclosed:

1) Declaration of John C. Rockett, Ph.D., under 37 C.F.R. § 1.132, with Exhibits A - Q;
2) [First] Declaration of Tod Bedilion, Ph.D., under 37 C.F.R. § 1.132, with Exhibits A - H;
3) [Second] Declaration of Tod Bedilion, Ph.D., under 37 C.F.R. § 1.132;
4) Declaration of Vishwanath R. Iyer, Ph.D., under 37 C.F.R. § 1.132 with Exhibits A - E; and
5) Ten (10) references published before the filing date of the instant application:
   a) WO 95/21944, SmithKline Beecham, "Differentially expressed genes in healthy and diseased subjects" (Aug. 17, 1995)
   b) WO 95/20681, Incyte Pharmaceuticals, "Comparative Gene Transcript Analysis" (Aug 3, 1995)
   c) Schena et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science 270:467-470 (Oct 20, 1995)
   d) WO 95/35505, Stanford University, "Method and apparatus for fabricating microarrays of biological samples" (Dec 28, 1995)
   e) U.S. Pat. No. 5,569,588, Ashby et al., "Methods for Drug Screening" (Oct 29, 1996)
   f) Heller al., "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays," *PNAS* 94:2150 - 2155 (Mar 1997)
   g) WO 97/13877, Lynx Therapeutics, "Measurement of Gene Expression Profiles in Toxicity Determinations" (April 17, 1997)
   h) Acacia Biosciences Press Release (August 11, 1997)
   i) Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," Genetic Engineering News (Sept. 15, 1997)
   j) DeRisi *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science 278:680 - 686 (Oct 24, 1997)

Reference 1 of 11
of Response dated 12/04/03
In USSN 09/937,060

Functional characterization of human sphingosine kinase-1

Victor E. Nava[a,1], Emanuela Lacana’[a,1], Samantha Poulton[a], Hong Liu[a], Masako Sugiura[b], Keita Kono[b], Sheldon Milstien[c], Takafumi Kohama[b,1], Sarah Spiegel[a,1,]*

[a] *Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, 353 Basic Science Building, 3900 Reservoir Road NW, Washington, DC 20007, USA*
[b] *Pharmacology and Molecular Biology Research Laboratories, Sankyo Co., Ltd., Tokyo 140-8710, Japan*
[c] *LCMR, NIMH, Bethesda, MD 20892, USA*

Received 9 March 2000; received in revised form 10 April 2000

Edited by Guido Tettamanti

Abstract Sphingosine kinase catalyzes the phosphorylation of sphingosine to form sphingosine 1-phosphate (SPP), a novel lipid mediator with both intra- and extracellular functions. Based on sequence identity to murine sphingosine kinase (mSPHK1a), we cloned and characterized the first human sphingosine kinase (hSPHK1). The open reading frame of hSPHK1 encodes a 384 amino acid protein with 85% identity and 92% similarity to mSPHK1a at the amino acid level. Similar to mSPHK1a, when HEK293 cells were transfected with hSPHK1, there were marked increases in sphingosine kinase activity resulting in elevated SPP levels. hSPHK1 also specifically phosphorylated D-*erythro*-sphingosine and to a lesser extent sphinganine, but not other lipids, such as D,L-*threo*-dihydrosphingosine, *N,N*-dimethylsphingosine, diacylglycerol, ceramide, or phosphatidylinositol. Northern analysis revealed that hSPHK1 was widely expressed with highest levels in adult liver, kidney, heart and skeletal muscle. Thus, hSPHK1 belongs to a highly conserved unique lipid kinase family that regulates diverse biological functions.

Key words: Human sphingosine kinase; Sphingosine 1-phosphate

1. Introduction

The metabolic product of sphingosine kinase (SPHK), sphingosine 1-phosphate (SPP), is a lipid signaling molecule that acts both intra- and extracellularly to affect many biological processes. These include mitogenesis [1,2], apoptosis [3], atherosclerosis [4] and inflammatory responses [5,6]. Specific members of the EDG-1 family of G protein-coupled receptors bind SPP (reviewed in [7,8]) and modulate chemotaxis [9,10], angiogenesis [10–12], neurite retraction and cell rounding [13]. Because SPP levels are mainly regulated by the activity of SPHK, cloning and characterization of this enzyme are important for understanding its role in normal and pathological processes. Previously, we purified SPHK to homogeneity from rat kidneys [14] and subsequently identified mouse cDNAs encoding two forms of SPHK, designated mSPHK1a and mSPHK1b, whose predicted proteins differ by only 10 amino acids at their N-terminus [15]. The corresponding mRNAs may arise by alternative splicing. In this study, sequence homologies to the mSPHK1a cDNAs were used to identify and clone the first human homologue, hSPHK1. hSPHK1 is ubiquitously expressed in adult tissues with highest levels in liver, kidney, lung and skeletal muscle. Our results suggest that hSPHK1 belongs to a family of highly conserved enzymes which differ from other known lipid kinases.

2. Materials and methods

2.1. Materials

SPP, sphingosine, and *N,N*-dimethylsphingosine (DMS) were from Biomol Research Laboratory Inc. (Plymouth Meeting, PA). All other lipids were purchased from Avanti Polar Lipids (Birmingham, AL). [γ-^{32}P]ATP (3000 Ci/mmol) was purchased from Amersham (Arlington Heights, IL). Poly-L-lysine was from Boehringer Mannheim (Indianapolis, IN). Alkaline phosphatase from bovine intestinal mucosa, type VII-NT, was from Sigma (St. Louis, MO). Restriction enzymes were from New England Biolabs (Beverly, MA). Lipofectamine Plus was from Life Technologies (Gaithersburg, MD).

2.2. Human sphingosine kinase cDNA cloning

BLAST searches using mSPHK1a sequences identified an EST clone (AA026479) which contained sequences homologous to several conserved domains of mSPHK [15]. To obtain a full-length cDNA, the 5′-end of hSPHK1 was extended by rapid amplification of cDNA ends/polymerase chain reaction (RACE-PCR; Life Technologies). First, cDNA was synthesized from HEK293 poly(A)$^+$ RNA with a gene-specific antisense primer hspk1-GSP1 (5′-ACCATTGTCCAGTGAG). Then two consecutive PCR reactions using LA Taq (TaKaRa) were performed. First PCR: 5′RACE Abridged Anchor Primer and the antisense primer hspk1-GSP2 (5′-TTCCTACAGGGAGGTAGGCC) at 94°C for 2 min followed by 30 cycles of amplification (94°C for 1 min, 55°C for 1 min, 72°C for 2 min) and primer extension at 72°C for 5 min. Second PCR: Abridged Universal Amplification Primer and the antisense primer hspk1-GSP3 (5′-GGCTGCCAGACGCAGGAAGG) using a program similar to the first PCR but with annealing at 65°C. The PCR products were cloned into pCR 2.1 (TA Cloning, Invitrogen) and sequences confirmed by automated sequencing. To make expression constructs, a primer set was designed as follows: sense primer containing a Kozak sequence and ATG start codon, sphk1-GSP4 (5′-GCCACCATGGATCCAGCGGGCGGCCCC); antisense primer, sphk1-GSP5 (5′-TCATAAGGGCTCTTCTGGCGGTGGCATCTG). The PCR reaction was performed using human fetus Marathon-Ready cDNA (Clontech) as template with the above primers, and the amplification product was subcloned into pCR3.1 (Eukaryotic TA Cloning, Invitrogen). In addition, hSPHK1 was tagged at the N-terminus by subcloning into a pcDNA-c-myc vector [2] using high fidelity taq polymerase (Pfu, Stratagene). hSPHK1 accession number is AF238083.

*Corresponding author. Fax: (1)-202-687 0260.
E-mail: spiegel@bc.georgetown.edu

[1] These authors contributed equally to this report.

Abbreviations: BSA, bovine serum albumin; DMS, *N,N*-dimethylsphingosine; DHS, D,L-*threo*-dihydrosphingosine; SPHK, sphingosine kinase; SPP, sphingosine 1-phosphate; PCR, polymerase chain reaction; RACE, rapid amplification of cDNA ends

0014-5793/00/$20.00 © 2000 Federation of European Biochemical Societies. All rights reserved.
PII: S0014-5793(00)01510-6

2.3. Cell culture and expression of sphingosine kinase

Human embryonic kidney cells (HEK293, ATCC CRL-1573) were grown in high glucose Dulbecco's modified Eagle's medium (DMEM) containing 100 U/ml penicillin, 100 μg/ml streptomycin and 2 mM L-glutamine supplemented with 10% fetal bovine serum [15]. Cells were transfected with either pcDNA3.1 or pCR3.1 containing hSPHK1 using Lipofectamine Plus according to the manufacturer's protocol. Transfection efficiencies were typically about 40%.

2.4. Measurement of sphingosine kinase activity

Cytosolic sphingosine kinase activity was determined with 50 μM sphingosine, dissolved in 5% Triton X-100 (final concentration 0.25%), and [γ-^{32}P]ATP (10 μCi, 1 mM) containing $MgCl_2$ (10 mM) as previously described [15]. In some experiments, sphingosine was added as a complex with bovine serum albumin (BSA) as previously described [15]. Specific activity is expressed as pmol SPP formed per min per mg protein.

2.5. Lipid extraction and measurement of SPP, sphingosine, and ceramide

Cells were washed with phosphate buffered saline and scraped in 1 ml of methanol containing 2.5 μl concentrated HCl. Lipids were extracted by adding 2 ml chloroform/1 M NaCl (1:1, v/v) and 100 μl 3 N NaOH and phases were separated. The basic aqueous phase containing SPP, and devoid of sphingosine, ceramide, and the majority of phospholipids, was transferred to a siliconized glass tube. The organic phases were re-extracted with 1 ml methanol/1 M NaCl (1:1, v/v) plus 50 μl 3 N NaOH, and the aqueous fractions combined. Mass measurements of SPP in the aqueous phase were carried out as previously described [16]. Sphingosine and ceramide in the organic phase were determined by enzymatic methods using sphingosine kinase and diacylglycerol kinase, respectively [17]. Total phospholipids present in lipid extracts were also quantified [17].

2.6. Northern blotting analysis

Poly(A)$^+$ RNA blots containing 2 μg of poly(A)$^+$ RNA per lane from multiple adult human tissues (Clontech) were hybridized with the 0.6 kb *Eco*RV/*Sph*I fragment of pCR3.1-hSPHK1, which was gel-purified and labeled with [^{32}P]dCTP by random priming. Hybridization in ExpressHyb buffer (Clontech) was carried out at 65°C overnight according to the manufacturer's protocol. Blots were reprobed with a human β-actin control probe (Clontech). Bands were quantified using a Molecular Dynamics Phosphoimager.

3. Results and discussion

3.1. Cloning of hSPHK1

BLAST searches of the EST database identified a human homologue of murine SPHK, EST AA026479, with similarity to the 3′ end of mSPHK1a. This sequence was used to design specific primers and 5′ RACE was performed on mRNA extracted from HEK293 cells to obtain the full-length cDNA of hSPHK1. The open reading frame encodes a protein with 384 amino acids, and 85% identity and 92% similarity to mSPHK1a at the amino acid level (Fig. 1). We previously found by sequence alignment that SPHKs from mouse, yeast and *Caenorhabditis elegans* share several conserved blocks of amino acids [15]. Similarly, hSPHK1 contains these conserved regions (C1–C5, Fig. 1), including the invariant positively charged motif, GGKGK, in the C1 domain, which may be part of the ATP binding site of this novel class of lipid kinases.

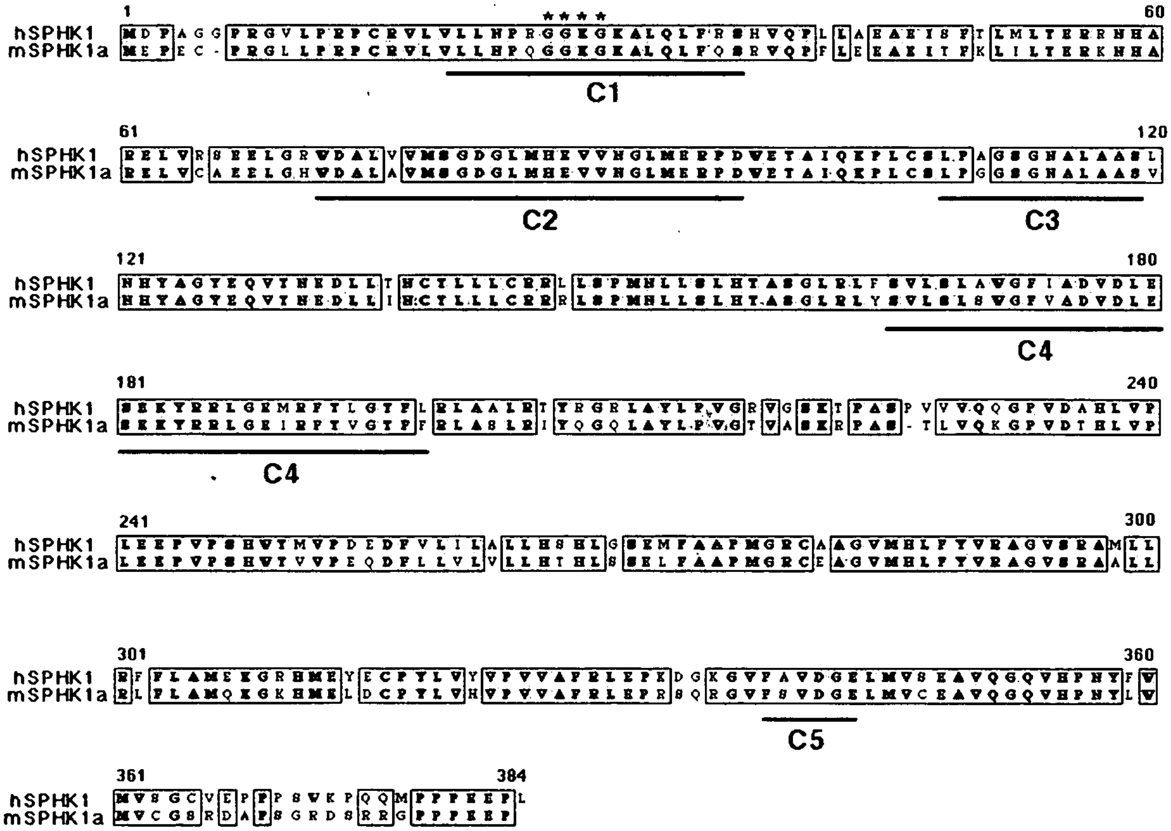

1
****
60
hSPHK1 MDPAGGPRGVLPRPCRVLVLLNPRGGKGKALQLFRSHVQPLLAEAEISFTLMLTERRNHA
mSPHK1a MEPEC-PRGLLPRPCRVLVLLNPQGGKGKALQLFQSRVQPFLEEAEITFKLILTERKNHA
C1
61
120
hSPHK1 RELVRSEELGRWDALVVMSGDGLMHEVVNGLMERPDWETAIQKPLCSLPAGSGNALAASL
mSPHK1a RELVCAEELGHWDALAVMSGDGLMHEVVNGLMERPDWETAIQKPLCSLPGGSGNALAASV
C2
C3
121
180
hSPHK1 NHYAGYEQVTNEDLLTNCTLLLCRRLLSPMNLLSLHTASGLRLFSVLSLAWGFIADVDLE
mSPHK1a NHYAGYEQVTNEDLLINCTLLLCRRRLSPMNLLSLHTASGLRLYSVLSLSWGFVADVDLE
C4
181
240
hSPHK1 SEKYRRLGEMRFTLGTFLRLAALRTYRGRLAYLPVGRVGSKTPASPVVVQQGPVDAHLVP
mSPHK1a SEKYRRLGEIRFTVGTFFRLASLRIYQGQLAYLPVGTVASKRPAS-TLVQKGPVDTHLVP
C4
241
300
hSPHK1 LEEPVPSHWTVMVPDEDFVLILALLHSHLGSEMFAAPMGRCAAGVMHLFYVRAGVSRAMLL
mSPHK1a LEEPVPSHWTVVPEQDFLLVLVLLHTHLSSELFAAPMGRCEAGVMHLFYVRAGVSRAALL
301
360
hSPHK1 RFFLAMEKGRHMEYECPYLVYVPVVAFRLEPKDGKGVFAVDGELMVSEAVQGQVHPNYFW
mSPHK1a RLFLAMQKGKHMELDCPYLVHVPVVAFRLEPRSQRGVFSVDGELMVCEAVQGQVHPNYLW
C5
361
384
hSPHK1 MVSGCVEPPPSWKPQQMPPPEEPL
mSPHK1a MVCGSRDAPSGRDSRRGPPPEEP

Fig. 1. Predicted amino acid sequence of hSPHK1 and alignment of the conserved domains. ClustalW alignment of SPHKs from mouse and human. Identical and conserved amino acid substitutions are shaded dark and light gray, respectively. The conserved domains (C1–C5) are indicated by lines and the invariant positively charged motif GGKGK by asterisks.

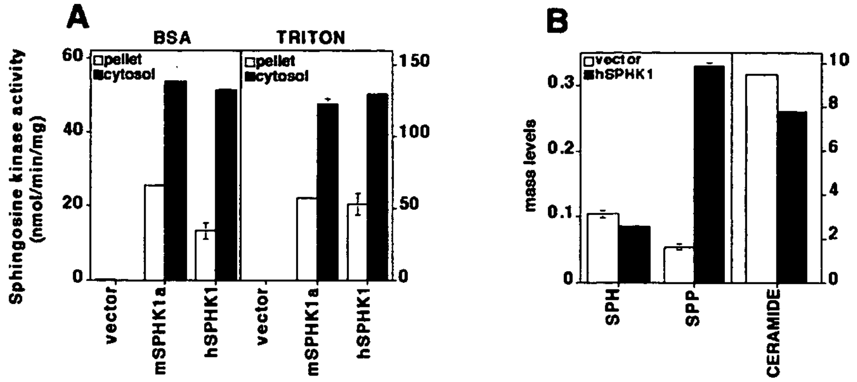


Fig. 2. Activity of hSPHK1 expressed in HEK293 cells. A: HEK293 cells were transiently transfected with empty vector or vector containing either mSPHK1a or hSPHK1. SPHK activity was measured in cytosol (filled bars) and particulate pellet (open bars) 24 h after transfection using sphingosine–BSA complexes or sphingosine–Triton X-100 micelles as substrate as indicated. SPHK activity in vector transfected cells was 84 ± 2 and 134 ± 27 pmol/min/mg using sphingosine–BSA complexes or sphingosine–Triton X-100 micelles as substrate, respectively. Data are means ± S.D. and are representative of two independent experiments performed in triplicate. B: Changes in mass levels of SPP, sphingosine, and ceramide. Mass levels of SPP, sphingosine and ceramide in cells transfected with empty vector (open bars) or vector containing hSPHK1 (filled bars) were measured after 24 h. Data are expressed as pmol/nmol phospholipid and are means ± S.D. of triplicate determinations.

3.2. *hSPHK1 encodes a functional sphingosine kinase*

HEK293 cells were transfected with expression vectors containing hSPHK1 to determine whether it encodes a bona fide SPHK. Modest levels of endogenous SPHK activity were detected in cells transfected with an empty vector (Fig. 2A). Twenty-four hours after transfection with pcDNA3.1-hSPHK1, the SPHK activity increased approximately 600-fold and remained at this level for at least 2 days. For comparison, a similar increase in activity was observed after transfection with mSPHK1a (Fig. 2A). Similar results were obtained when cells were transfected with hSPHK1 in pCR3.1. In agreement with previous results with mSPHK1a [15], hSPHK1 was stimulated by Triton X-100. Both membrane-associated and cytosolic SPHK activity have been described in mammalian tissues and cell lines [1,18–21]. In cells transfected with hSPHK1, approximately 70% of the SPHK activity was found in the cytosol and only about 30% was membrane-associated (Fig. 2A). Similarly, we previously found that the majority of mSPHK1a activity was also expressed in the cytosol [2,15]. Kyte–Doolittle hydropathy plots did not suggest the presence of any potential hydrophobic membrane spanning domains in the primary structure of hSPHK1.

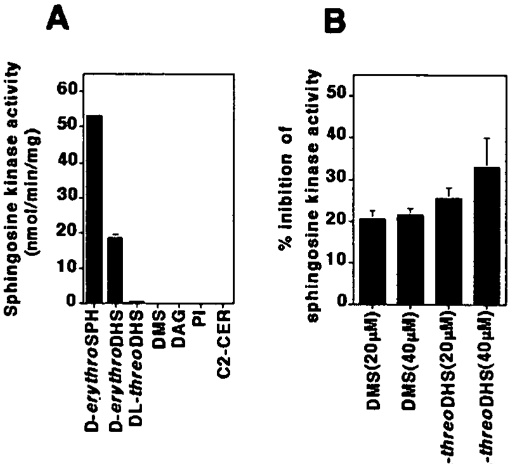


Fig. 3. A: Substrate specificity of hSPHK1. HEK293 cells were transfected with hSPHK1 and SPHK-dependent phosphorylation of various sphingosine analogs or other lipids (50 μM) was measured using cell lysates as enzyme source. DAG, diacylglycerol; PI, phosphatidylinositol; C2-CER, *N*-acetyl-sphingosine. B: DMS and DHS are inhibitors of hSPHK1. SPHK activity in HEK293 cell lysates 24 h after transfection with hSPHK1 was measured with 10 μM SPP in the absence or presence of 20 μM and 40 μM DMS or DHS. Data are means ± S.D. of triplicate determinations and are expressed as percent inhibition.

Transfection of HEK293 cells with hSPHK1 also resulted in changes in levels of sphingolipid metabolites (Fig. 2B). Mass levels of SPP increased 5.7-fold compared to cells transfected with vector alone, with a 18% decrease in levels of both sphingosine and ceramide. However, because intracellular ceramide pools are much larger than sphingosine pools, the absolute decrease of ceramide was greater than the decrease in sphingosine mass. These results suggest that transfected hSPHK1 is active in intact cells, and that kinase overexpression can alter the intracellular balance of sphingolipid metabolites.

3.3. *Substrate specificity of hSPHK1*

The naturally occurring D-(+)-*erythro*-*trans*-isomer of sphingosine and *erythro*-dihydrosphingosine (sphinganine) were the best substrates for hSPHK1 (Fig. 3A). However, similar to the specificity of mSPHK1a [15], sphingosine was more efficiently phosphorylated than sphinganine. Moreover, other sphingo-

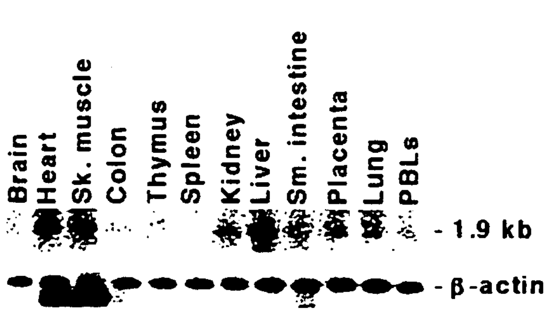


Fig. 4. Tissue-specific expression of hSPHK1 by Northern blot analysis. Top panel: A hSPHK1 probe was hybridized to a poly(A)$^+$ RNA blot with the human tissues indicated at the top of each lane as described in Section 2. Bottom panel: A β-actin probe was used to reprobe the blot.

lipids, including D,L-*threo*-dihydrosphingosine (DHS) and C2-ceramide, as well as diacylglycerol and phosphatidylinositol were not substrates (Fig. 3A). With D-*erythro*-sphingosine as substrate, half-maximal velocity was found at 5 μM, in excellent agreement with K_m values previously determined with rat kidney SPHK [14] and recombinant mSPHK1a [15]. DMS and DHS have previously been used to inhibit SPHK and block increases in SPP induced by various physiological stimuli [1,3,22]. Both of these sphingolipids also inhibited hSPHK1 and similar to their inhibitory effects on mSPHK1a [15], DHS was slightly more potent than DMS (Fig. 3B).

3.4. *Tissue distribution of hSPHK1 expression*

The tissue distribution of SPHK1 mRNA expression in adult human tissues was analyzed by Northern blotting (Fig. 4). In most tissues, including adult brain, heart, spleen, lung, kidney, and testis, a predominant 1.9 kb mRNA species was detected. Expression was highest in adult liver, heart and skeletal muscle. In comparison, we previously showed that mSPHK1a expression is greatest in mouse spleen, lung, kidney, testis and heart, with much lower expression in skeletal muscle [15].

In summary, hSPHK1 is the human homolog of mSPHK1. Based on EST sequences, hSPHK1 has been localized on chromosome 17q25.2 at the marker stSG28540 (D17S785-D17S836 Reference Interval, UniGene cluster Hs. 68061, URL: http://www.ncbi.nlm.nih.gov/unigene/clust.cgi?org = hs and cid = 68061). hSPHK1 belongs to a conserved family of genes that is distinct from other known lipid kinases. Molecular cloning and characterization of members of the SPHK family should help to clarify their potential roles in various human diseases as their product, SPP, has been implicated as an important regulatory component of biological processes including growth, survival, allergy, chemotaxis, and angiogenesis.

Acknowledgements: This work was supported by a grant from the National Institutes of Health GM43880 (to S.S.) and a Postdoctoral Fellowship BC961968 from the United States Army Medical Research and Materiel Command, Prostate Cancer Research Program (V.E.N.). We thank Ayako Yamamoto for excellent technical assistance and Dr. Lisa C. Edsall for measurements of ceramide.

References

[1] Olivera, A. and Spiegel, S. (1993) Nature 365, 557–560.

[2] Olivera, A., Kohama, T., Edsall, L.C., Nava, V., Cuvillier, O., Poulton, S. and Spiegel, S. (1999) J. Cell Biol. 147, 545–558.

[3] Cuvillier, O., Pirianov, G., Kleuser, B., Vanek, P.G., Coso, O.A., Gutkind, S. and Spiegel, S. (1996) Nature 381, 800–803.

[4] Xia, P., Vadas, M.A., Rye, K.A., Barter, P.J. and Gamble, J.R. (1999) J. Biol. Chem. 274, 33143–33147.

[5] Xia, P., Wang, L., Gamble, J.R. and Vadas, M.A. (1999) J. Biol. Chem. 274, 34499–34505.

[6] Prieschl, E.E., Csonga, R., Novotny, V., Kikuchi, G.E. and Baumruker, T. (1999) J. Exp. Med. 190, 1–8.

[7] Spiegel, S. (1999) J. Leukocyte Biol. 65, 341–344.

[8] Goetzl, E.J. and An, S. (1998) FASEB J. 12, 1589–1598.

[9] Sadahira, Y., Ruan, F., Hakomori, S. and Igarashi, Y. (1992) Proc. Natl. Acad. Sci. USA 89, 9686–9690.

[10] Wang, F., Van Brocklyn, J.R., Edsall, L., Nava, V.E. and Spiegel, S. (1999) Cancer Res. 59, 6185–6191.

[11] Lee, O.H., Kim, Y.M., Lee, Y.M., Moon, E.J., Lee, D.J., Kim, J.H., Kim, K.W. and Kwon, Y.G. (1999) Biochem. Biophys. Res. Commun. 264, 743–750.

[12] Lee, M.J., Lee, M.J., Thangada, S., Claffey, K.P., Ancellin, N., Liu, C.H., Kluk, M., Volpi, M., Sha'afi, R.I. and Hla, T. (1999) Cell 99, 301–312.

[13] Van Brocklyn, J.R., Tu, Z., Edsall, L.C., Schmidt, R.R. and Spiegel, S. (1999) J. Biol. Chem. 274, 4626–4632.

[14] Olivera, A., Kohama, T., Tu, Z., Milstien, S. and Spiegel, S. (1998) J. Biol. Chem. 273, 12576–12583.

[15] Kohama, T., Olivera, A., Edsall, L., Nagiec, M.M., Dickson, R. and Spiegel, S. (1998) J. Biol. Chem. 273, 23722–23728.

[16] Edsall, L.C. and Spiegel, S. (1999) Anal. Biochem. 272, 80–86.

[17] Edsall, L.C., Pirianov, G.G. and Spiegel, S. (1997) J. Neurosci. 17, 6952–6960.

[18] Stoffel, W., Heimann, G. and Hellenbroich, B. (1973) Hoppe-Seyler's Z. Physiol. Chem. 354, 562–566.

[19] Buehrer, B.M. and Bell, R.M. (1992) J. Biol. Chem. 267, 3154–3159.

[20] Olivera, A., Rosenthal, J. and Spiegel, S. (1994) Anal. Biochem. 223, 306–312.

[21] Ghosh, T.K., Bian, J. and Gill, D.L. (1994) J. Biol. Chem. 269, 22628–22635.

[22] Choi, O.H., Kim, J.-H. and Kinet, J.-P. (1996) Nature 380, 634–636.

MOLECULAR CARCINOGENESIS 24:153–159 (1999)

IN PERSPECTIVE
Claudi J. C nti, Editor

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1] *Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina*
[2] *Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland*

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3′ regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnostics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

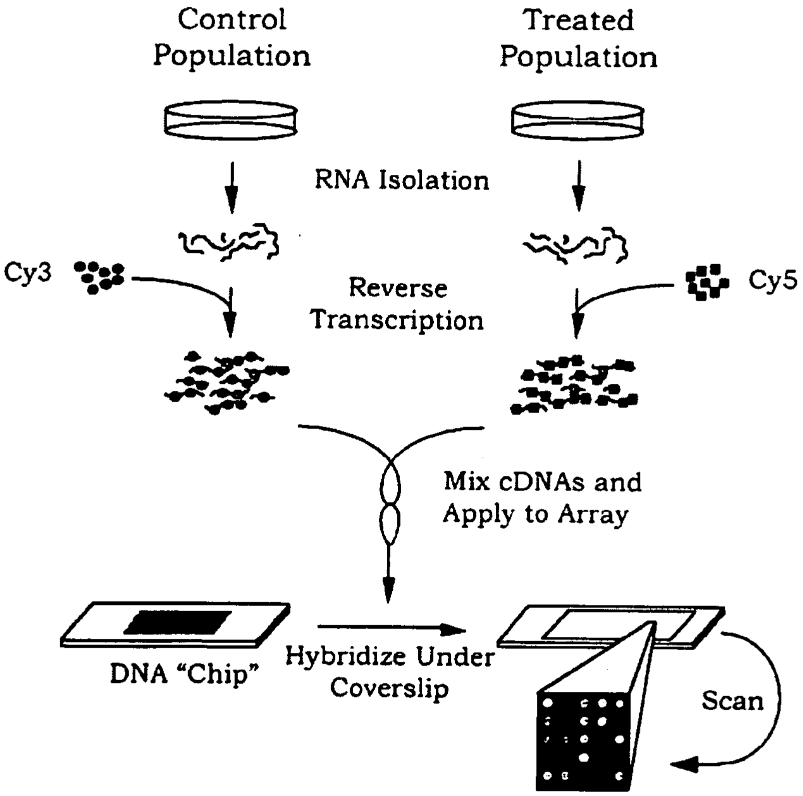


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

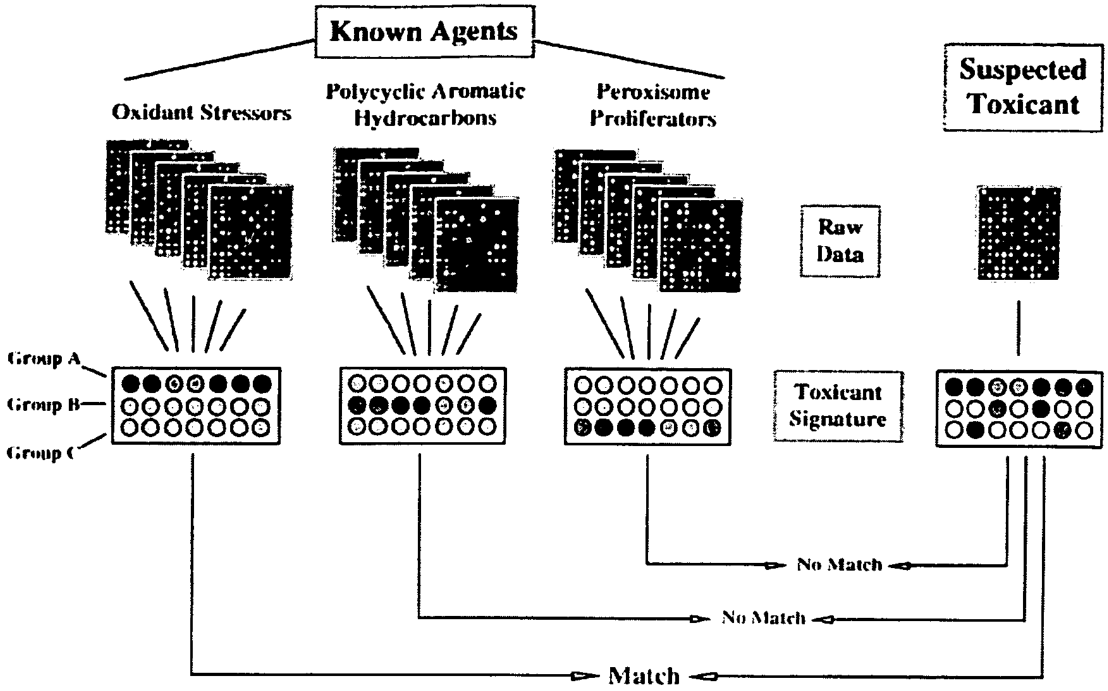


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip D signed to Det ct R sponses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker,Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

ELSEVIER

Toxicology Letters 112–113 (2000) 467–471

T xicol gy
Letters

www.elsevier.com/locate/toxlet

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: + 1-301-4245989; fax: + 1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

0378-4274/00/$ - see front matter © 2000 Elsevier Science Ireland Ltd. All rights reserved.
PII: S0378-4274(99)00236-2

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

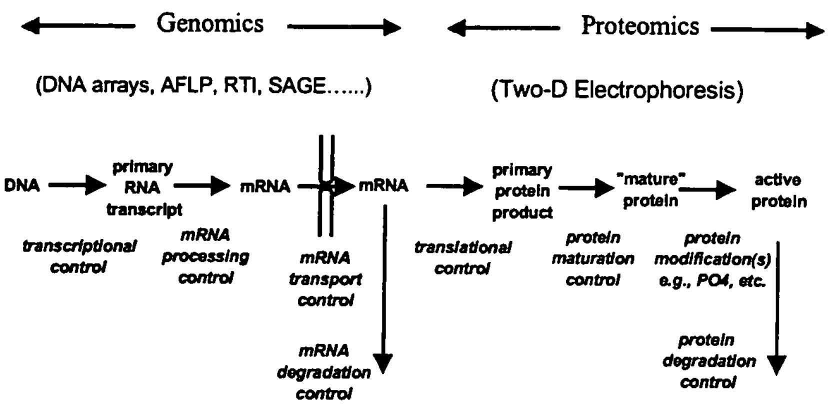


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

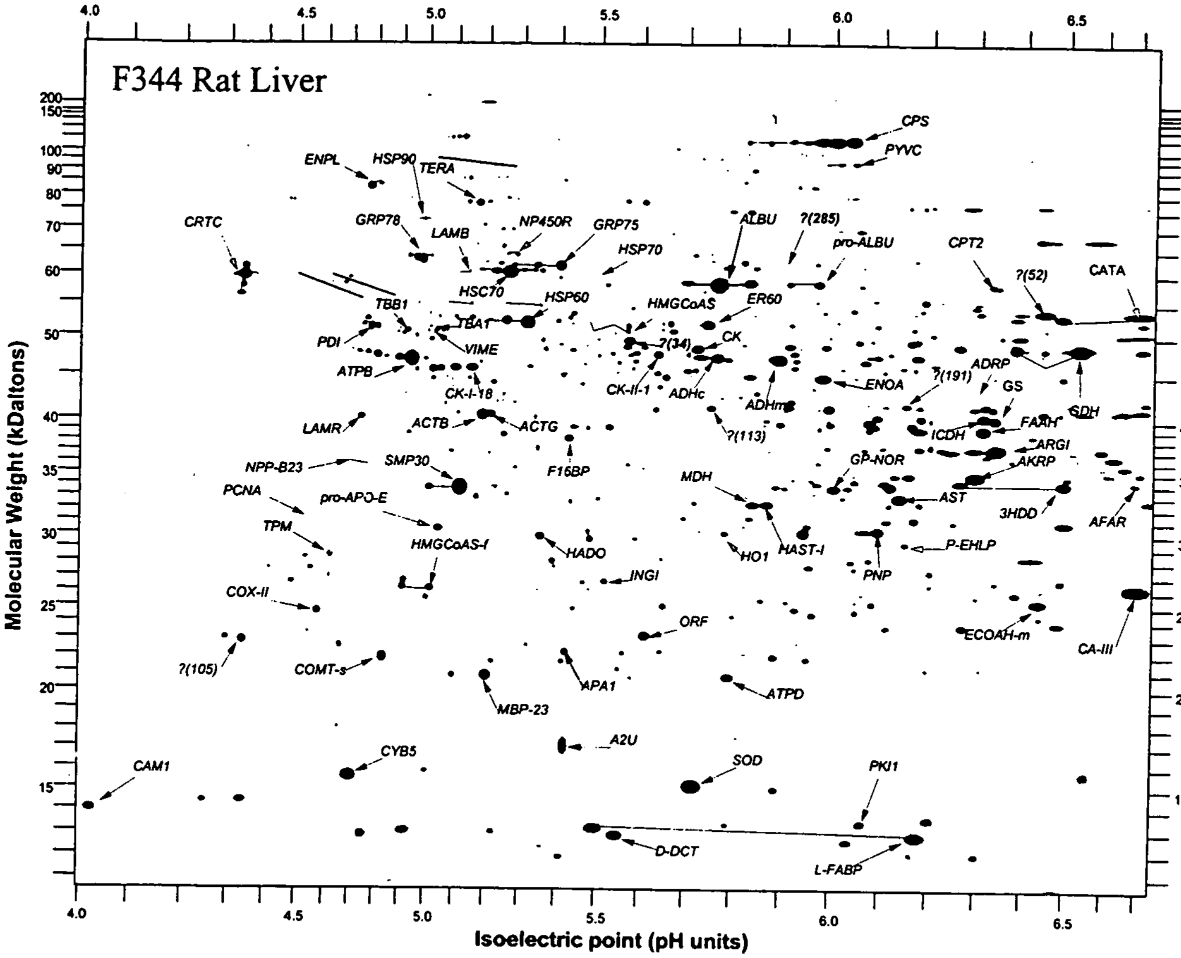


Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplification. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug development, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepsdorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expresssion patterns with a complementary DNA microarray. Science 251, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Rayrnackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Envir nmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, N rth Car lina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations.

Key words: DNA arrays, gene arrays, microarrays, toxicology. ***Environ Health Perspect*** 107:681–685 (1999). [Online 6 July 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA r protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of tw types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (*3*). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm^2 may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ^{33}P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (*4*). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

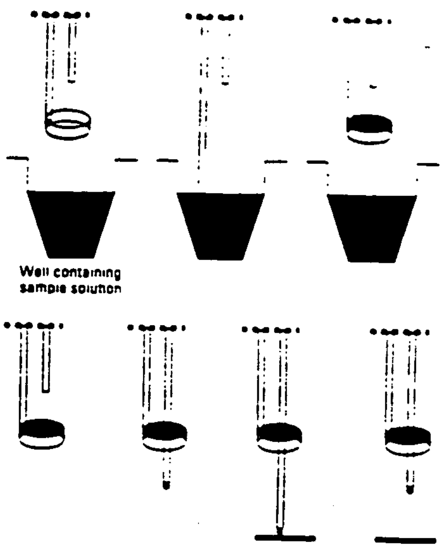


Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (*14*), with permission from Genetic Microsystems.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reacti ns may also be used. Fluorophore- or radi labeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5′ end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case flu r-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series f labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Table 1. Advantages and disadvantages of different microarray scanning systems.

	CCD camera system	Nonconfocal laser scanner	Confocal laser scanner
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples	—	May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (*13*).

Analysis of DNA Micr arrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire gen me f the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (*5*). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (*6,7*). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C. elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is ften confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (*8*) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previ usly characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings. Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

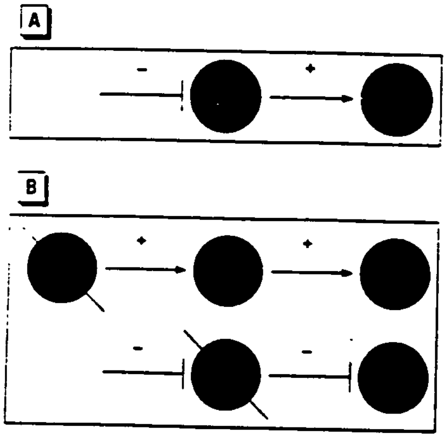


Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (*A*) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and j_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (*B*) deleting j_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether j_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (*15*).

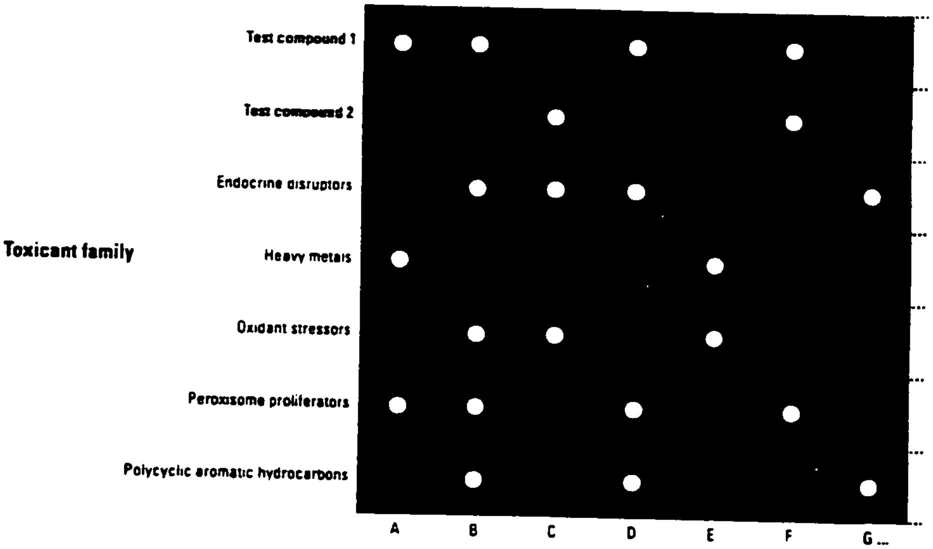


Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (*9*), and how this may compromise sperm counts and quality following sexual maturation (*10*). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (*11*) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (*12*):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

Speakers

Cindy Afshari
NIEHS
Linda Birnbaum
U.S. EPA
Ron Butow
University of Texas
Southwestern Medical Center
Alex Chenchik
Clontech Laboratories, Inc.
David Dix
U.S. EPA
Abdel Elkahloun
Research Genetics, Inc.
Sue Fenton
U.S. EPA
Norman Hecht
University of Pennsylvania
Pat Hurban
Paradigm Genetics, Inc.
Bob Kavlock
U.S. EPA
Ernie Kawasaki
General Scanning, Inc.
Steve Krawetz
Wayne State University
Nick Mace
Genetic Microsystems, Inc.
Scott Mordecai
Affymetrix, Inc.
Kevin Morgan
Glaxo Wellcome, Inc.
Elaine Poplin
Research Genetics, Inc.
Don Rose
Cartesian Technologies, Inc.
Jim Samet
U.S. EPA
Sam Ward
University of Arizona
Jeff Welch
U.S. EPA
Reen Wu
University of California at Davis
Tim Zacharewski
Michigan State University

References and Notes

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:56–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overheck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

Diacylglyceride kinases, sphingosine kinases and NAD kinas s: distant r latives of 6-phosphofructokinases

Gilles Labesse, Dominique Douguet, Liliane Assairi and Anne-Marie Gilles

Diacylglyceride kinases, sphingosine kinases, NAD kinases and 6-phosphofructokinases are thought to be related despite large evolution of their sequ nces. Discovery of a common signature has led to the suggestion that they possess a similar phosphate-donor-binding site and a similar phosphorylation mechanism. The substrate- and allosteric-binding sites are much more divergent and their delineation remains to be determined xperimentally.

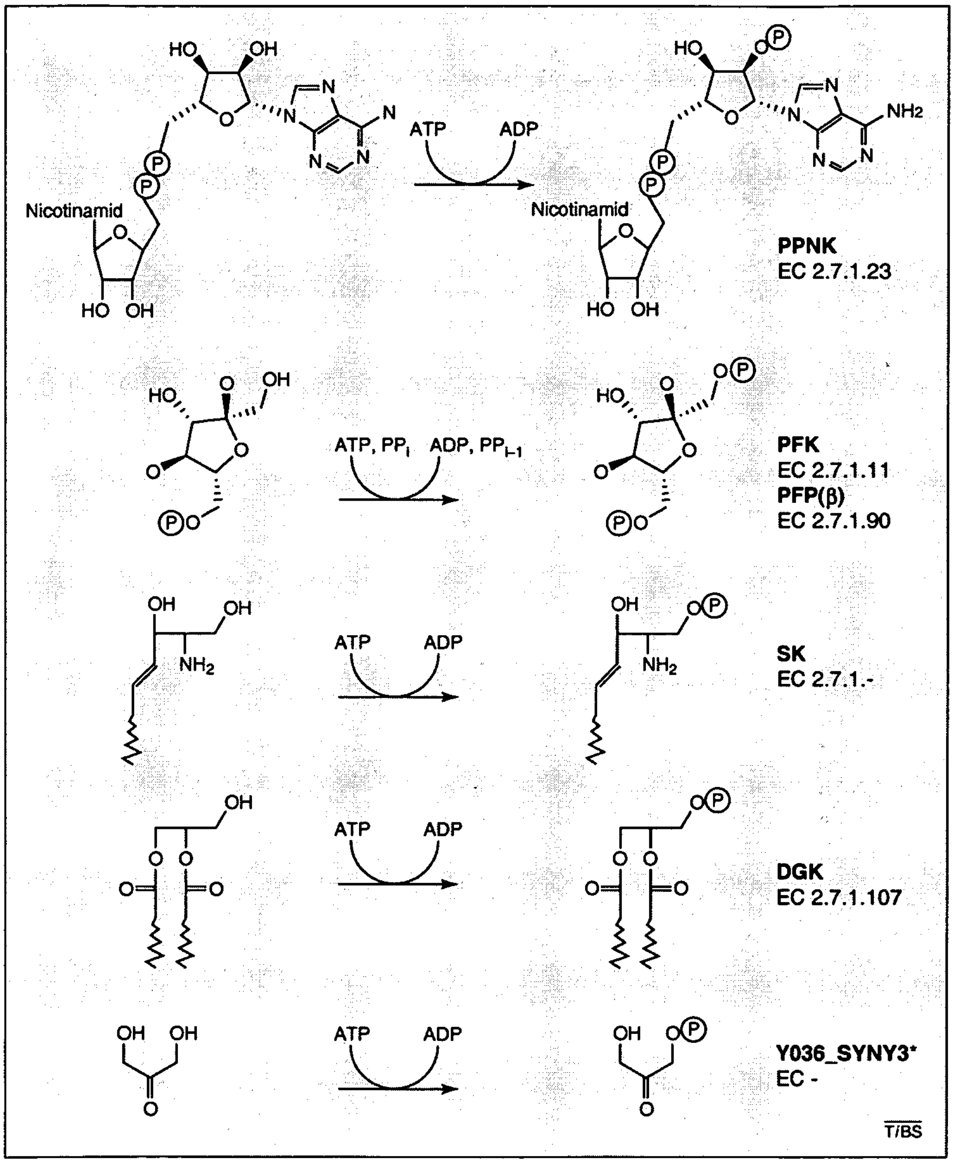


Fig. 1. Reactions catalyzed by DGKs, SKs, PFKs, PPNKs and possibly Y036_SYNY3. The phosphate group transferred to the product is circled. PP$_i$ indicates the polyphosphate molecules of various length used by PFPs; ATP is the preferred phosphate donor for the other kinases. Long acyl chains are shown as zigzag lines in the structure of the substrate of SK and DGK. *: The putative function of Y036_SYNY3 was deduced by similarity and its putative substrate specificity was predicted from domain-swapping analysis. Abbreviations: DGKs, diacylglyceride kinases; PFKs, 6-phosphofructokinases; PFP(β), pyrophosphate-dependant phosphofructokinase (β subunit); PPNKs, polyphosphate/ATP NAD kinases; SKs, sphingosine kinases.

Despite their importance in key metabolic and signaling pathways, diacylglyceride kinases (DGKs), sphingosine kinases (SKs) and polyphosphate/ATP NAD kinases (PPNKs) are poorly characterized enzymes. At present, there are no structural data available for these kinases. By contrast, the 6-phosphofructokinases (PFKs) are well-characterized enzymes that have no protein relatives other than the pyrophosphate-dependent phosphofructokinases (PFPs). Extending sequence comparisons of all these kinases to structural analysis, threading and sequence motif searches, suggested that they are weakly related, both with respect to structure and function. These observations explain some of the experimental data discussed in the following text (e.g. results from directed mutagenesis), and also suggest a function for a related hypothetical protein (SWISSPROT: Y036_SYNY3).

PFKs (EC 2.7.1.11; Pfam PF00365) phosphorylate D-fructose 6-phosphate (Fig. 1) and regulate glycolytic flux. The crystal structure of this allosteric enzyme was solved in complex with its substrate and a phosphate donor [1]. PFK is ~300 amino acids (aa) in length, and comprises two similar (α/β) lobes: one involved in ATP binding and the other housing both the substrate-binding site and the allosteric site (a regulatory binding site distinct from the active site, but that affects enzyme activity). Both PFKs and the related PFPs [2] (EC 2.7.1.90) are dimeric or tetrameric. In the crystal structure [1], the phosphate donor (ADP) binds a specific sequence motif, GGdGs (upper and lower case letters refer to strictly and loosely conserved amino acids, respectively), which contains the aspartate involved in Mg^{2+} chelation.

The phosphorylation of NAD was recently shown to be catalyzed by a member of the PPNK family [polyphosphate/ATP–NAD kinase,

http://tibs.trends.com 0968-0004/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0968-0004(02)02093-5

```
SK sub-family:
JPREDSPH2                  bbbbbb        aaaaaaaaa aaaa    bbbbbb                bbbbb    bbbbbb
SPH2_HUMAN     (139)-LLPRPPRLLLLVNPFGGRGLAWQWCKNHVLPMISEAGLSFNLIQTER-(11)-SLSEWDGIVTVSGDGLLHEVLNGLLDRP-(07)-

DGK sub-family:
JPREDY036                  bbbbbb        aaaaaaaaaaaa       bbb                bbbbbb     aaaaaaaa
Y036_SYNY3     (120)-LLGKTKTGHLIFNPVAGQGNVERELDLIKEHLQSEINLKITFTSAEV-(20)-DGEGDSPIIASGGDGTVSGVAAALVNTG-(00)-
BMRU_BACSU       (0)---MSHRKALLIHNGNAANKNIEKALGAVVPVLSHHLDEVIIKQTKKK-(10)--DDSVDTVFILGGDGTIHQCINAILERK-(00)-
KDGZ_HUMAN     (475)-PTPSPKPLLVFVNPKSGGNQGAKIIQSFLWYLNP--RQVFDLSQGGP-(00)-----NLRILACGGDGTVGWILSTLDQLR-(04)-
JPREDKDGZ                  bbbbb       aaaaaaaaaaaa       bbbb                 bbbbbb    aaaaaaaaaaaa

PPNK sub-family:
JPREDUTR1                 bbbbbb     bbaaaaaaaaaaaaa     bbbbbbaaaa           bbbbbbb    bbbaaaaaa
UTR1_YEAST     (122)-VELDVENLMIVTKLNDVSLYFLTRELVEWVLVHF--RVTVYVDSELK-(31)-HDVFFDLVVTLGGDGTVLFVSSIFQRHV-(00)-
PPNK_ECOLI       (0)-MNNHFKCIGIVGHPRHPTALTTHEMLYRWLCTKG---YEVIVEQQIA-(17)---QLADLAVVVGGDGNMLGAARTLARYD-(00)-
PPNK_HELPY       (0)-MKDSLQTIGVFVRPTHYQNPLFEKLEQAKEWVL-----KLLEDEGFE-(14)-LIEKADAFLCLGGDGTILGALRMTHSYN-(00)-
JPREDPPNK                 bbbbbb     aaaaaaaaaaaaaaa      bbbb                 bbbbbb    aaaaaaaaa

PFK sub-family:
K6PF_HUMAN      (14)------KAIAVLTSGGDAQGMNAAVRAVVRVGIPTGARVFFVHEGYQG-(48)-VKRGITNLCVIGGDGSLTGADTFRSEWS-(22)-
1PFK             (0)----MIKKIGVLTSGGDAPGMNAAIRGVVRSALTEGLEVMGIYDGYLG-(46)-KKRGIDALVVIGGDGSYMGAMRLTEMG--(00)-
P-SEA                     bbbbbbbbbbbb aaaaaaaaaaaaaaa bbbbbbb  aaa       aaa    bbbbbbbaaaaaaaaaaaa
                                                                              *************************

SK sub-family:
JPREDSPH2         bb      aaaaa       bbbb             bbbbbb             bbbbbb                 bbbb
SPH2_HUMAN    MPVGILPCGSGNALAGA-(88)-GRLSYLPATVE(152)-DFVLMLAISPS-(16)-GLVHLCWVRSGIPS-(36)-LTPRGVLTVDGE-(29)

DGK sub-family:
JPREDY036       bbbb         aa      bbb    aaaa      bbbbbbbb            bbbbbb                 bbb
Y036_SYNY3    IPLGIIPRGTANAFSVA-(33)-LLAGVGFEAEM-(44)-EASAITIANAA-(17)-GLLDITVASSQTAL-(35)-TSPPQKIVVDGE-(26)
BMRU_BACSU    PAVGILPGGTSNDFSRV-(33)-NFWGIGLIAET-(44)-EAVMLLVMNGQ-(16)-GLLDVLICRNTNLT-(22)-TDTAKKADMDGE-(24)
KDGZ_HUMAN    PPVAILPLGTGNDLART-(58)-NYFSLGFDAHV-(60)-KPQCVVFLNIP-(25)-GYLEVIGFTMTSLA-(19)-TSKAIPVQVDGE-(326)
JPREDKDGZ       bbbb        bbb          b b aaa         bbbbbb         bbbbbbb aaaaa             bbb

PPNK sub-family:
JPREDUTR1       bbbb       bb          bbbbb              bbbb             bbbbb                bbbbb
UTR1_YEAST    PPVMSFSLGSLGFLTNF-(55)-ILNEVTIDRGP-(19)-QADGLIAATPT-(15)-PTVNAIALTPICPH-(21)-KSRPAWAAFDGK-(126)
PPNK_ECOLI    IKVIGINRGNLGFLTDL-(30)-AINEVVLHPGK-(19)-RSDGLIISTPT-(15)-PSLDAITLVPMFPH-(21)-RRNDLEISCDSQ-(45)
PPNK_HELPY    KPCFGVRIGNLGFLSAV-(35)-AINEIVIAKKK-(19)-KGDGLIIATPL-(15)-ALSQSYILTPLCDF-(21)-AHEDALVVIDGQ-(49)
JPREDPPNK       bbbb       bb        aabbbbbb             bbbb             bbbbb                bbbbb

PFK sub-family:
K6PF_HUMAN    LNIVGLVGSIDNDFCGT-(30)-FVLEVMGRHCG-(10)-GADWVFIPECP-(19)-GSRLNIIIVAEGAI-(21)-YDTRVTVLGHVQ-(482)
1PFK          FPCIGLPGTIDNDIKGT-(38)-SVVEVMGRYCG-(10)-GCEFVVVPEVE-(16)-GKKHAIVAITEHMC-(14)-RETRATVLGHIQ-(71)
P-SEA         bbbbbb                 bbbbbb              bbbb  bb          bbbbbbbb           bbbbbbb
              ****************
```

Fig. 2. Multiple sequence alignment of DGKs, PFKs, PPNKs and related enzymes in the region of the conserved motifs. The alignment was performed manually. Sequence codes are from the SWISSPROT database [4] and the corresponding secondary structure predictions, obtained from JPRED2 [17], were renamed accordingly (e.g. JPREDY036 for Y036_SYNY3). The secondary structure assignment for the crystal structure PDB1PFK [1] was performed using P-SEA (http://bioserv.cbs.cnrs.fr/HTML_BIO/frame_sea.html). Predicted α helices and β strands are denoted by a purple 'a' and a red 'b', respectively. The Asp residue in the conserved motif 'gGdgs' is highlighted with a yellow background and the motifs previously defined for the identification of DGKs and SKs are underlined. The position of the signature derived using PATTINPROT is shown by asterisks. Following the PATTINPROT [23] nomenclature, its sequence reads as: X(40)-{WEKG}-{RWVI}-[AGNTVIMLFYHR]-[ASVILFM]-[CVILFYHM]-[GAPCSTVILFM]-[GASCVILM]-[SG]-G-[EDN]-[GDN]-[STFLIVAEDN]-[ATVIMLFYHRN]-X(9,35)-{TWCFM}-[PCVILMFKT]-[GASTPINHQKR]-[ASCTVIFLM]-[GASCVIMLFY]-[GASCTPVIF]-[ACTVIMLFRNG]-[GASNHRKPV]-[GATDEKRYLIMVC]-[GSDTI]-[GASTPVILFNHR]-[GASNDVIMLF]-[GNTVW]-[APTDLF]-[CVIMFYWLK]-{WEQHFL}. 'X' stands for any amino acid, numbers in parentheses indicate the number of possible repetitions, [] surrounds possible substitutions at one position and {} corresponds to unallowed amino acids. Abbreviations: DGKs, diacylglyceride kinases; PFKs, 6-phosphofructokinases; PPNKs, polyphosphate/ATP NAD kinases; SKs, sphingosine kinases. This multiple sequence alignment (alignment number ALIGN_000335) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000335.dat).

EC 2.7.1.23; Fig. 1]. Members of this family are 260–320 aa in length and are usually dimeric or tetrameric. The recently identified PPNK from *Mycobacterium tuberculosis* [3] (SWISSPROT [4]: PPNK_MYCTU), and orthologs from distinct complete genomes, were used for PSI–BLAST searches [5]; these revealed weak sequence similarities (E-value range: 1.00–0.02) to various PFKs.

DGKs (EC 2.7.1.107) phosphorylate diacylglycerol (Fig. 1), which is a second messenger that activates protein kinase C and is important in cell regulation [6]. The monomeric mammalian isoenzymes (550–1170 aa) possess several domains [6,7]. For example, the C terminus houses the catalytic domain, which is regulated by anionic amphiphiles (e.g. phosphatidylserine). DGKs also possess an original motif (named DAGKc in SMART [8] and PF00781 in Pfam [9]) that contains the sequence 'ϕϕϕGGDGT' (ϕ stands for any hydrophobic residue). Eukaryotic DGKs are related to as-yet-uncharacterized bacterial homologs, including Y036_SYNY3.

The monomeric SKs (~49 kDa) are related to DGKs. They phosphorylate sphingosine to form sphingosine 1-phosphate (Fig. 1), which acts both as an intracellular second messenger (e.g. in the inhibition of apoptosis) and as a ligand for a family of G-protein-coupled receptors [10]. SKs are regulated by acidic phospholipids (e.g. dioleoylphosphatidylserine) [10,11]. Several regions are highly conserved among SKs, including the motif 'ϕϕϕϕSGDGi'. Mutation of the second glycine in this SK signature (G82D in human SK) abolishes the kinase activity [11], as does mutation of the equivalent glycine residue in several DGKs [7,11]. SKs and DGKs are novel kinases sharing a common ~350-aa-long catalytic domain; they have no significant similarity to other known kinases [7].

Using PSI–BLAST with default parameters to search the SWISSPROT database with Y036_SYNY3 (aa 1–433) as a query revealed significant similarities, at convergence (14 iterations), between the C terminus of the query and: (1) DGKs (E-value ranging from e^{-95} to e^{-60}) and (2) PPNKs (e^{-60} to e^{-27}). The N terminus (aa 1–116) of the query showed significant similarities with bacterial methylglyoxal synthases [12]. These enzymes use dihydroxyacetone-phosphate (DHA-P) as a substrate, thereby suggesting that the kinase domain of the query might be involved in DHA phosphorylation (Fig. 1). Similarities to the PFKs appeared just below the default threshold of 0.002. These sequence similarities can be extended to PFKs by adding a couple of PFK sequences (E-value: 0.02–0.50) to the inclusion set before resuming the PSI–BLAST search. At the second convergence, 101 sequences from the SWISSPROT database are detected. Searches with a slightly higher threshold (0.006 and 0.004 instead of 0.002) gave the same final results in only one convergence (20th and 15th iteration, respectively). Shifting to the more complete and non-redundant database GenPept, and using the matrices PAM70 (inclusion threshold: 0.002) or BLOSUM80 (inclusion threshold: 0.004), confirmed the previous results. At convergence, after 21 iterations (PAM70), DGKs, SKs, PFKs and PPNKs showed significant sequence similarities (e^{-52} to e^{-6}), which extended over a common region of ~300 aa.

The fold compatibility between these enzymes was further analyzed using 3D-PSSM [13], FUGUE [14], GenTHREADER [15], PDB–BLAST (http://bioinformatics.burnham-inst.org/pdb_blast/), SAM-T99 [16] and J-PRED2 [17] through our meta-server [18]. With most queries (including those in Fig. 2), PDB–BLAST and SAM-T99, used with default parameters, showed weak but significant similarities (E-values <0.01) to PFK. Fold recognition results were further analyzed for single-domain protein sequences to avoid any noise resulting from possible domain–domain interfaces and improper delimitation of the domains. The best scores (corresponding to >95% certainty) were obtained using the sequence PPNK_ECOLI (3D-PSSM, E-value: $8.16e^{-02}$ and FUGUE, Z-score: 5.27). The other observed hits corresponded mostly to related (α/β) folds.

Analysis of the resulting multiple alignment deduced from the PSI–BLAST search showed that DGKs, PPNKs and PFKs align in the same common region, forming a short, well-conserved motif (Fig. 2). A PHI–BLAST 2.1.2 search [19] using the various queries and the short seed pattern [GS]-G-[ED]-G-[ST] also revealed significant similarities (~0.004) to the Pfam profile of the PFKs (PF00365).

A PATTINPROT search [20] was performed starting with the motif 'ϕdϕϕϕϕϕgGdgs' to refine the signature of the common region. The pattern was extended to a unique and specific signature that is common to these four previously unrelated kinase subfamilies (finding 230 sequences of DGKs, SKs, PPNKs and PFKs in the non-redundant database and 87 sequences out of 101 in SWISSPROT, the latter set containing partial sequences and inactive PFP α subunits). This new motif showed a highly significant E-value of $2.7e^{-11}$. This signature encompassed both the ATP- and substrate-binding sites of the crystal structure of PFK [1], and also comprised the two surrounding hydrophobic strands (highlighted by asterisks in Fig. 2). This signature also contained the previously described short motifs specific to the DGKs and SKs (underlined in Fig. 2), and explained the results of directed mutagenesis obtained on DGKs [7,21,22] and SKs [11]. Similarly, directed mutagenesis of the common motif also inactivated PPNK (L. Assairi and A-M. Gilles, unpublished). Its conservation suggested that these kinases might possess a similar ATP-binding site and might catalyze the phosphorylation using a common and specific mechanism.

These results suggest that these kinases would belong to the same superfamily and might adopt the PFK fold despite the very low overall sequence identity (10–20% over ~250 aa; Fig. 2) [23]. The deduced alignment should help us design new experiments to characterize these kinases; for example, to precisely delineate their specific substrate-binding site. Directed mutagenesis is currently undertaken on PPNKs to define the NAD-recognition motif.

References

1 Shirakihara, Y. and Evans, P.R. (1988) Crystal structure of the complex of phosphofructokinase from *Escherichia coli* with its reaction products. *J. Mol. Biol.* 204, 973–994

2 Carlisle, S.M. *et al.* (1990) Pyrophosphate-dependent phosphofructokinase. Conservation of protein sequence between the α- and β-subunits and with the ATP-dependent phosphofructokinase. *J. Biol. Chem.* 265, 18366–18371

3 Kawai, S. *et al.* (2000) Inorganic polyphosphate/ATP-NAD kinase of *Micrococcus flavus* and *Mycobacterium tuberculosis* H37Rv. *Biochem. Biophys. Res. Commun.* 276, 57–63

4 Bairoch, A. and Apweiler, R. (2000) The SWISS–PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48

5 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI–BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402

6 Kanoh, H. *et al.* (1990) Diacylglycerol kinase: a key modulator of signal transduction? *Trends Biochem. Sci.* 15, 47–50

7 Topham, M.K. and Prescott, S.M. (1999) Mammalian diacylglycerol kinases, a family of lipid kinases with signaling functions. *J. Biol. Chem.* 274, 11447–11450

8 Schultz, J. *et al.* (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.* 95, 5857–5864

9 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266

10 Pyne, S. and Pyne, N.J. (2000) Sphingosine 1-phosphate signalling in mammalian cells. *Biochem. J.* 349, 385–402

11 Pitson, S.M. *et al.* (2000) Expression of a catalytically inactive sphingosine kinase mutant blocks agonist-induced sphingosine kinase activation. A dominant-negative sphingosine kinase. *J. Biol. Chem.* 275, 33945–33950

12 Saadat, D. and Harrison, D.H. (1999) The crystal structure of methylglyoxal synthase from *Escherichia coli*. *Struct. Fold. Des.* 7, 309–317

13 Kelley, L.A. *et al.* (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299, 499–520

14 Shi, J. *et al.* (2001) FUGUE: sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310, 243–257

15 McGuffin, L.J. *et al.* (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16, 404–405

16 Karplus, K. *et al.* (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14, 846–856

17 Cuff, J.A. *et al.* (1998) JPRED: a consensus secondary structure prediction server. *Bioinformatics* 14, 892–893

18 Douguet, D. and Labesse, G. (2001) Easier threading through web-based comparisons and cross-validations. *Bioinformatics* 17, 752–753

19 Zhang, Z. *et al.* (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26, 3986–3990

20 Combet, C. *et al.* (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.* 25, 147–150

21 Sakane, F. *et al.* (1996) The C-terminal part of diacylglycerol kinase α lacking zinc fingers serves as a catalytic domain. *Biochem. J.* 318, 583–590

22 Masai, I. *et al.* (1993) *Drosophila* retinal degeneration A gene encodes an eye-specific diacylglycerol kinase with cysteine-rich zinc-finger motifs and ankyrin repeats. *Proc. Natl. Acad. Sci. U. S. A.* 90, 11157–11161

23 Labesse, G. (1996) MulBlast 1.0: a multiple alignment of BLAST output to boost protein sequence similarity analysis. *Comput. Appl. Biosci.* 12, 463–467

Gilles Labesse*
Dominique Douguet
Centre de Biochimie Structurale, INSERM
U554 – CNRS UMR5048, Universite
Montpellier I, 15, Av. Charles Flahault,
34060 Montpellier Cedex, France.
*e-mail: labesse@cbs.cnrs.fr

Liliane Assairi
Anne-Marie Gilles
Laboratoire de Chimie Structurale des
Macromolécules, CNRS URA 2185, Institut
Pasteur, 28 rue du Dr Roux, 75724 Paris,
Cedex 15, France.

Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should produce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic[1,2] and microarray[3,4] technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

Martin J. Page*, **B b Am ss**, **Christian Rohlff**, **C lin Stubb rfield**and **Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk

1359-6446/99/$ – see front matter © Elsevier Science. All rights reserved. PII: S1359-6446(98)01291-4

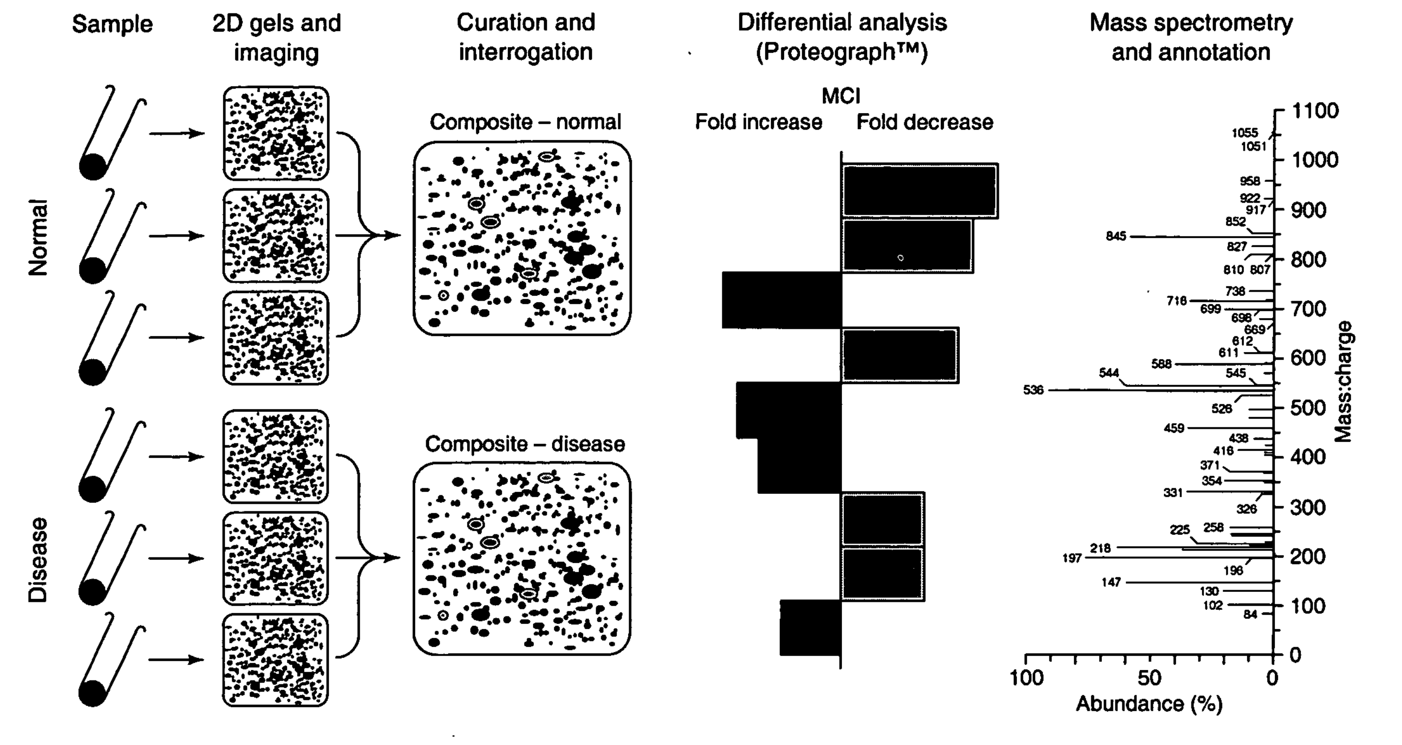


***Figure 1.** Steps involved in analysing a biological sample by proteomics. MCI, molecular cluster index.*

analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

Pr teomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed[5–7]. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins[8], which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS–PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

Us f proteomics to identify dis as specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and immediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical

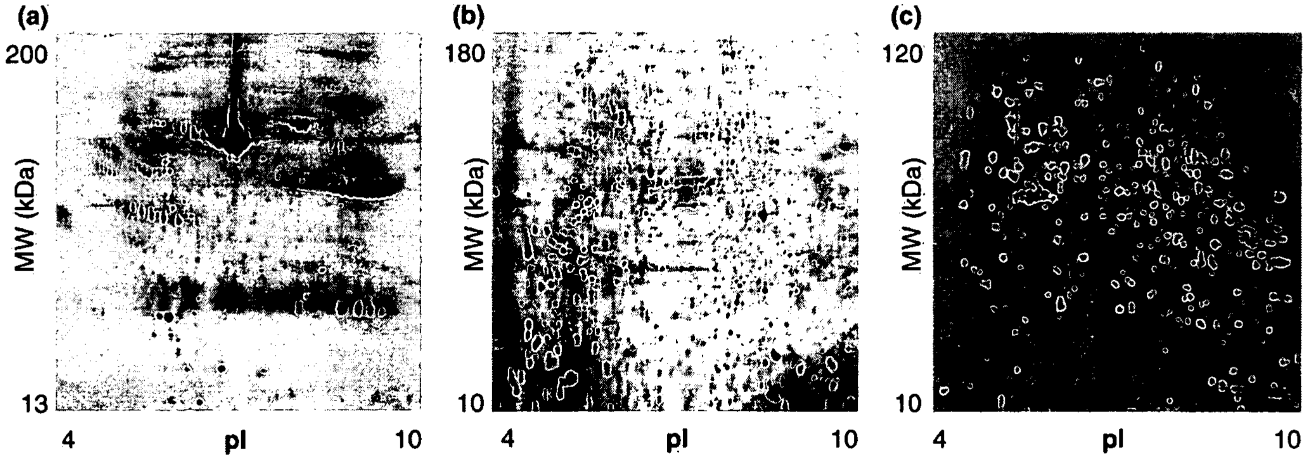
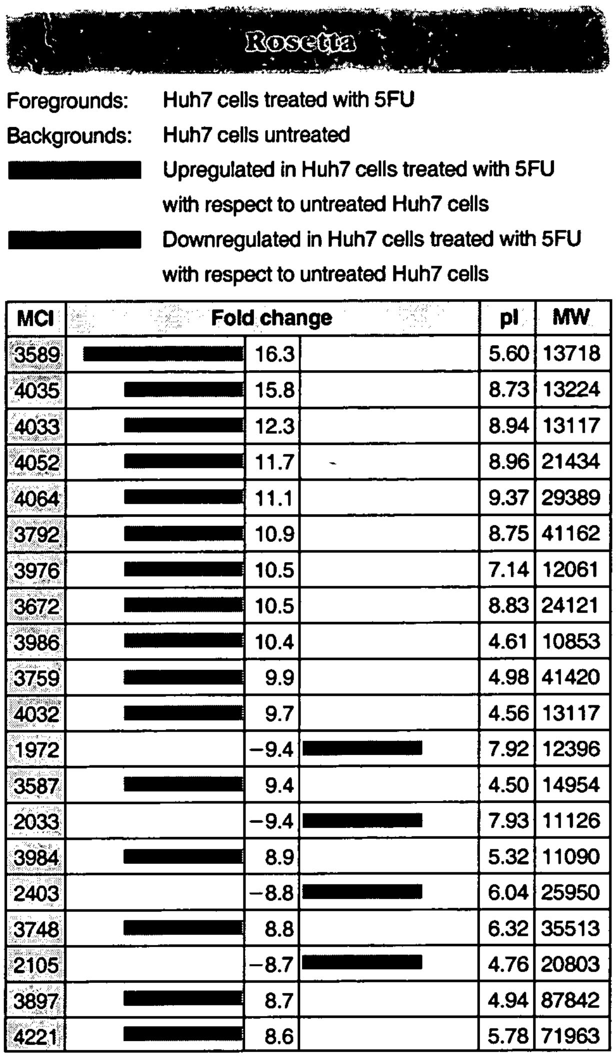


Figure 2. *Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus* Candida albicans *and (c) the human hepatoma cell line Huh7.*



Rosetta

Foregrounds: Huh7 cells treated with 5FU

Backgrounds: Huh7 cells untreated

█ Upregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

█ Downregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

MCI	Fold change			pI	MW
3589	█	16.3		5.60	13718
4035	█	15.8		8.73	13224
4033	█	12.3		8.94	13117
4052	█	11.7		8.96	21434
4064	█	11.1		9.37	29389
3792	█	10.9		8.75	41162
3976	█	10.5		7.14	12061
3672	█	10.5		8.83	24121
3986	█	10.4		4.61	10853
3759	█	9.9		4.98	41420
4032	█	9.7		4.56	13117
1972		−9.4	█	7.92	12396
3587	█	9.4		4.50	14954
2033		−9.4	█	7.93	11126
3984	█	8.9		5.32	11090
2403		−8.8	█	6.04	25950
3748	█	8.8		6.32	35513
2105		−8.7	█	4.76	20803
3897	█	8.7		4.94	87842
4221	█	8.6		5.78	71963

Figure 3. *Table of differential protein expression profiles, referred to as a Rosetta Proteograph™, between Huh 7 cells with and without the cytotoxic agent 5-FU. Bars are quantized and do not represent exact fold change values.*

cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a twofold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry[9]. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas[10], human breast proteins from normal and tumour sources[11–13], lung tumours[14], colon tumours[15] and bladder tumours[16]. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified[17,18].

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

Prot mics for targ t validati n and signal transducti n studi s

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences[20].

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules[21–23]. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics[24,25]. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affinity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies[26–28]. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable

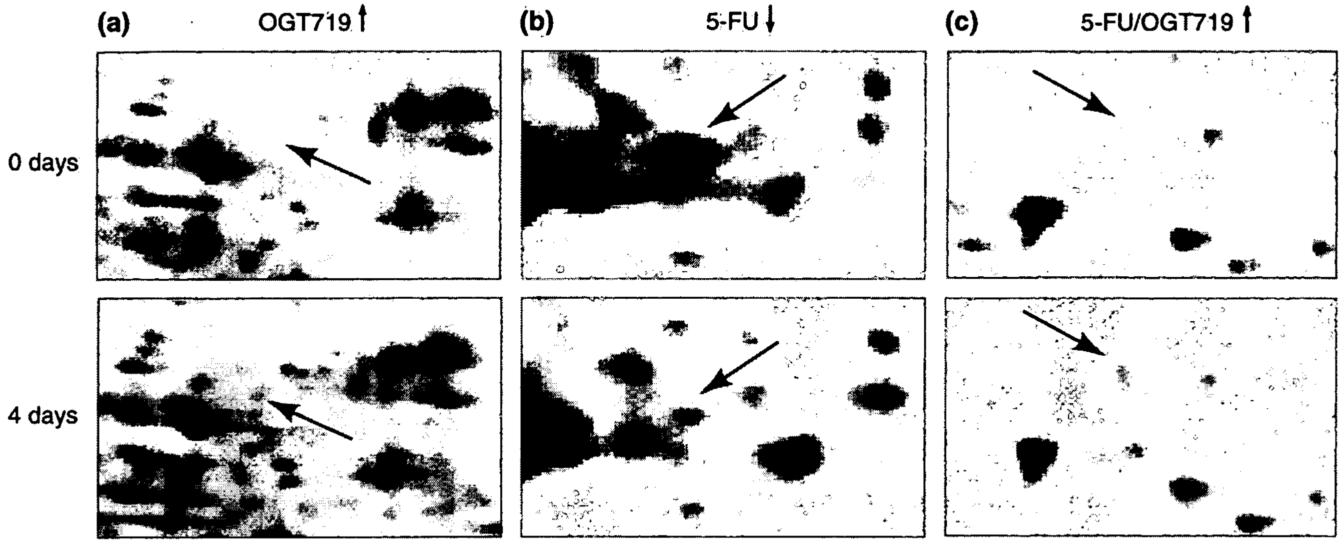


Figure 4. *Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1α2, (b) novel (three peptides by MS-MS) and (c) α-subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.*

of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes[29], hepatoma Huh7 cells[30] and some colorectal tumour cells[31]. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC_{50} doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins. Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein[32], can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

Clear potential

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

Us f proteomics in formal drug t xicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members[33,34], encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabolism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

Unique P450 profiles

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up[35–37]. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

REFERENCES

1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30

2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695

3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306

4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44

5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861

6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6

7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50

8 Parekh, R.B. and Rohlff, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723

9 Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818

10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751

11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581

12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343

13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825

14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848

15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613

16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117

17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794

18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18

19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537

20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783

21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159

22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359

23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641

24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091

25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880

26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277

27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125

28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323

29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229

30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330

31 Mu, J-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491

32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575

33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628

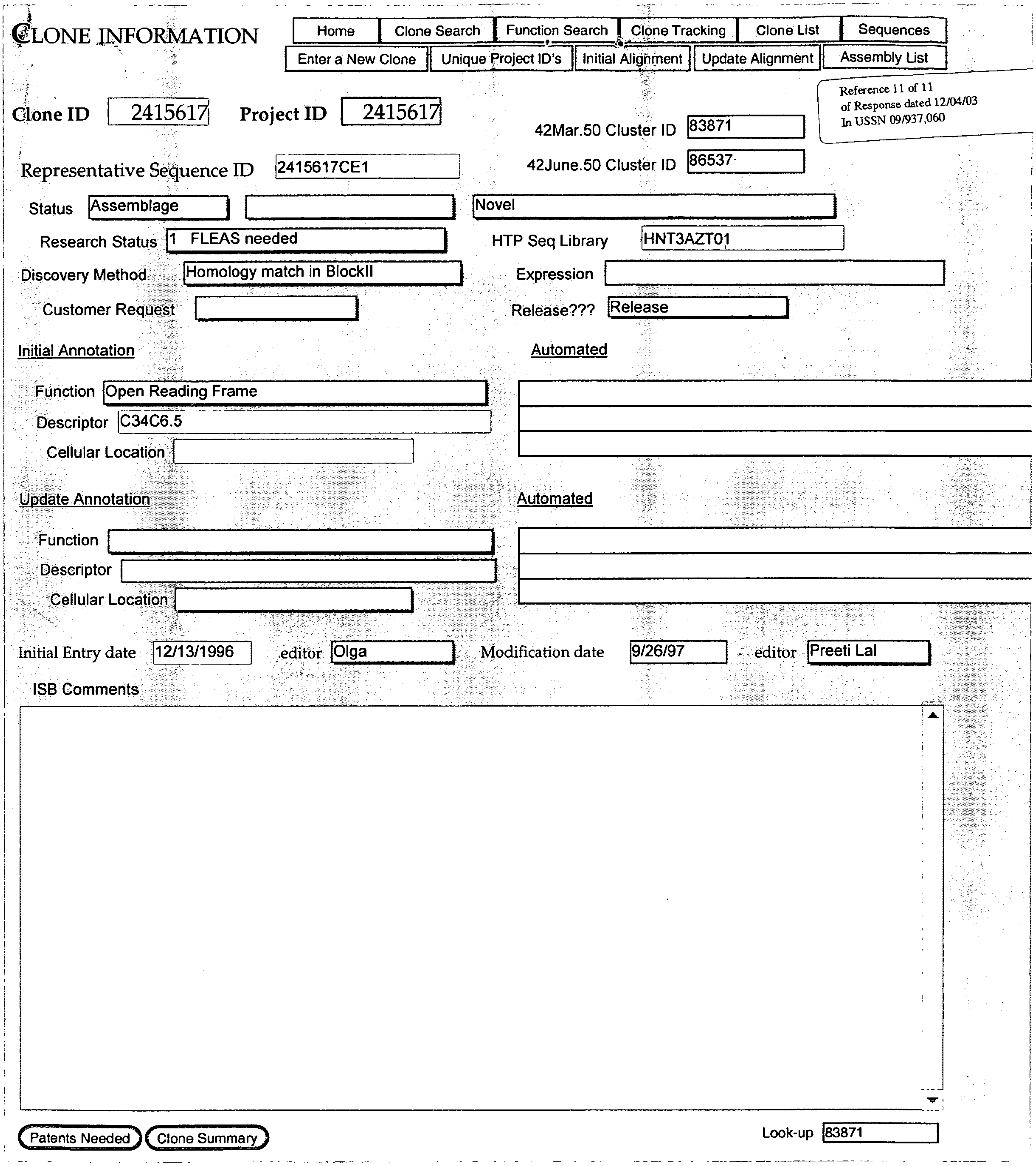
34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580

35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339

36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493

37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210

Reference 11 of 11
of Response dated 12/04/03
In USSN 09/937,060



CLONE INFORMATION

Clone ID 2415617 Project ID 2415617

42Mar.50 Cluster ID 83871

42June.50 Cluster ID 86537

Representative Sequence ID 2415617CE1

Status Assemblage Novel

Research Status 1 FLEAS needed

HTP Seq Library HNT3AZT01

Discovery Method Homology match in BlockII

Expression

Customer Request

Release??? Release

Initial Annotation

Automated

Function Open Reading Frame

Descriptor C34C6.5

Cellular Location

Update Annotation

Automated

Function

Descriptor

Cellular Location

Initial Entry date 12/13/1996 editor Olga Modification date 9/26/97 editor Preeti Lal

ISB Comments

Look-up 83871

Exhibit P of J. C. Rockett Declaration
dated 12/04/03
In USSN 09/937,060

US005807755A

United States Patent [19]

Ekins

[11] **Patent Number:** **5,807,755**

[45] **Date of Patent:** ***Sep. 15, 1998**

[54] **DETERMINATION OF AMBIENT CONCENTRATIONS OF SEVERAL ANALYTES**

[75] Inventor: **Roger P. Ekins**, London, Great Britain

[73] Assignee: **Multilyte Limited**, Great Britain

[*] Notice: The term of this patent shall not extend beyond the expiration date of Pat. No. 5,432,099.

[21] Appl. No.: **447,820**

[22] Filed: **May 23, 1995**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 984,264, Dec. 1, 1992, Pat. No. 5,432,099, which is a continuation of Ser. No. 460,878, filed as PCT/GB88/00649 Aug. 5, 1988, abandoned.

[30] **Foreign Application Priority Data**

Feb. 10, 1998 [GB] United Kingdom 8803000

[51] **Int. Cl.**^{6} **G01N 33/543**; G01N 33/537; G01N 33/533

[52] **U.S. Cl.** **436/518**; 436/501; 436/517; 435/7.1; 435/7.92; 435/973

[58] **Field of Search** 435/973, 7.1, 7.92; 436/518, 517, 501

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,385,126	5/1983	Chen et al.	436/518
5,432,099	7/1995	Ekins	436/518
5,486,452	1/1996	Gordon et al.	435/5

FOREIGN PATENT DOCUMENTS

8401031	3/1984	WIPO	G01N 33/54
8801058	2/1988	WIPO	G01N 33/543

OTHER PUBLICATIONS

White et al., "An Evaluation of Confocal Versus Conventional Imaging . . .," J Cell Biol 105: 41–48 (1987).
Ekins et al., "Development of Microspot Multi-Analyte Ratiometric . . .," Anal Chim Acta 227: 73–96 (1989).

Primary Examiner—Michael P. Woodward
Attorney, Agent, or Firm—Dann, Dorfman, Herrell and Skillman

[57] **ABSTRACT**

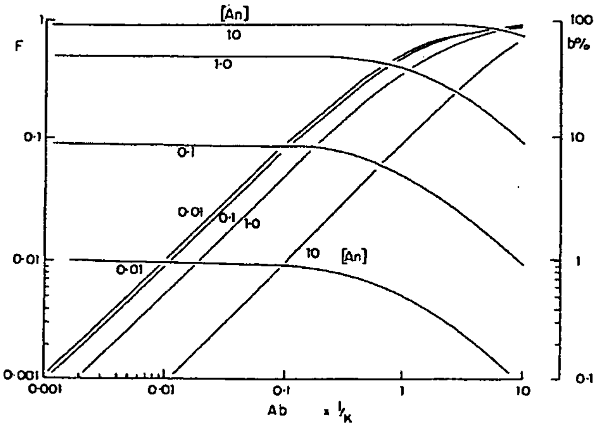
A method for determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprises

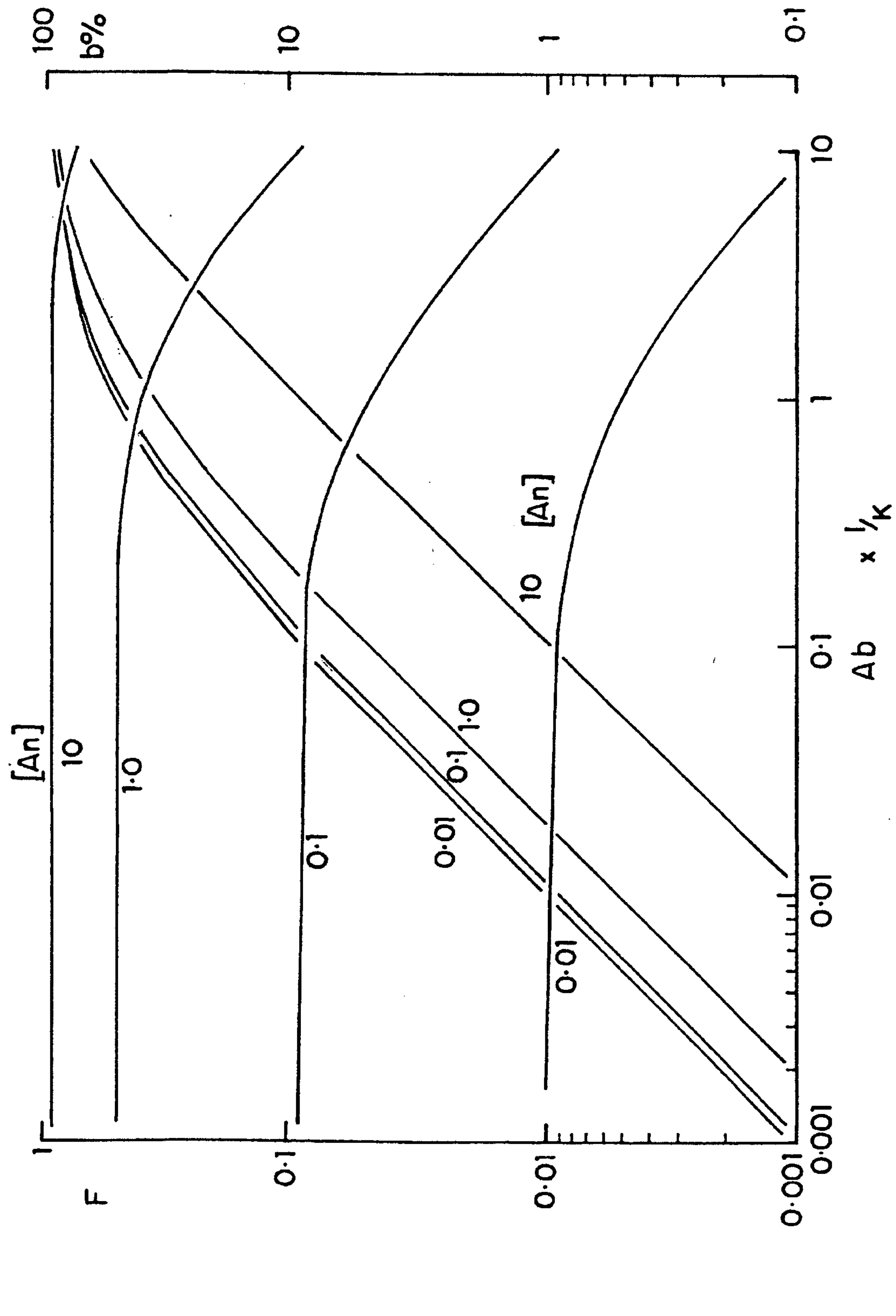
loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for that analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart locations such that each location has not more than 0.1 V/K, preferably less than 0.01 V/K, moles of a single binding agent, where K liters/mole is the equilibrium constant of the binding agent for the analyte;

contacting the loaded support means with the liquid sample to be analyzed, such that each of the spaced apart locations is contacted in the same operation with the liquid sample, the amount of liquid used in the sample being such that only an insignificant proportion of any analyte present in the liquid sample becomes bound to the binding agent specific for it, and

measuring a parameter representative of the fractional occupancy by the analytes of the binding agents at the spaced apart locations by a competitive or non-competitive assay technique using a site-recognition reagent for each binding agent capable of recognizing either the unfilled binding sites or the filled binding sites on the binding agent, said site-recognition reagent being labelled with a marker enabling the amount of said reagent in the particular location to be measured. A device and kit for use in the method are also provided.

17 Claims, 1 Drawing Sheet




F
1
0·1
0·01
0·001
b%
100
10
1
0·1
0·001
0·01
0·1
1
10
Ab x $^{1}/_{K}$
[An]
10
1·0
0·1
0·01
[An]
10
1·0
0·1
0·01

DETERMINATION OF AMBIENT CONCENTRATIONS OF SEVERAL ANALYTES

This application is a continuation-in-part of U.S. patent application Ser. No. 07/984,264, filed Dec. 1, 1992, now U.S. Pat. No. 5,432,099, which is a continuation of U.S. patent application Ser. No. 07/460,878, filed Feb. 2, 1990, now abandoned, filed as PCT/GB88/00649, Aug. 5, 1988.

FIELD OF THE INVENTION

The present invention relates to the determination of ambient analyte concentrations in liquids, for example the determination of analytes such as hormones, proteins and other naturally occurring or artificially present substances in biological liquids such as body fluids.

BACKGROUND OF THE INVENTION

I have proposed in International Patent Application WO84/01031 to measure the concentration of an analyte in a fluid by contacting the fluid with a trace amount of a binding agent such as an antibody specific for the analyte in the sense that it reversibly binds the analyte but not other components of the fluid, determining a quantity representative of the proportional occupancy of binding sites on the binding agent and estimating from that quantity the analyte concentration. In that application I point out that, provided that the amount of binding agent is sufficiently low that its introduction into the fluid causes no significant diminution of the concentration of ambient (unbound) analyte, the fractional occupancy of the binding sites on the binding agent by the analyte is effectively independent of the absolute volume of the fluid and of the absolute amount of binding agent, i.e. independent within the limits of error usually associated with the measurement of fractional occupancy. In such circumstances, and in these circumstances only, the initial concentration [H] of analyte in the fluid is related to the fraction (Ab/Ab_o) of binding sites on the binding agent occupied by the analyte by the equation:

$$\frac{Ab}{Ab_o} = \frac{K_{ab}[H]^2}{1+K_{ab}[H]}$$

where K_{ab} (hereinafter referred to as K) is the equilibrium constant for the binding of the analyte to the binding sites and is a constant for a given analyte and binding agent at any one temperature. This constant is generally known as the affinity constant, especially when the binding agent is an antibody, for example a monoclonal antibody.

The concept of using only a trace amount of binding agent is contrary to generally recommended practice in the field of immunoassay and immunometric techniques. For example, in such a well-known work as "Methods in Investigative and Diagnostic Endocrinology", ed. S. A. Berson and R. S. Yalow, 1973 at pages 111–116, it is proposed that in the performance of a competitive immunoassay maximum sensitivity of the assay is achieved if the proportion of the "tracer" analyte that is bound approximates to 50%. In order to achieve such a high degree of binding of the analyte the theory of Berson and Yalow, to this day generally accepted by other workers in the field, requires that the concentration of binding agent (or, strictly speaking, of binding sites, each molecule of binding agent conventionally having one or at most two binding sites) must be greater than or equal to the reciprocal of the equilibrium constant (K) of the binding agent for the analyte, i.e. $[Ab] \geqq 1/K$. For a sample of volume V the total amount of binding agent (or binding sites) must therefore be greater than or equal to V/K. A binding agent which is a monoclonal antibody may, for example, have an equilibrium constant (K) which is of the order of 10^{11} liters/mole for the specific antigen to which it binds. Thus, under the above generally accepted practice, a binding agent (or site) concentration of the order of 10^{-1} mole/liter or more is required for binding agents of such an equilibrium constant and, with fluid sample volumes of the order of 1 milliliter, the use of 10^{-14} or more mole of binding agent (or site) is conventionally deemed necessary. Avogadro's number is about 6×10^{23} so that 10^{-14} mole of binding site is equivalent to more than 10^9 molecules of binding agent even assuming that the binding agent possesses two binding sites per molecule. For specific binding agents of the very highest affinity K is less than 10^{13} liters/mole so that conventional practice requires more than 10^7 molecules of binding agent, whereas binding agents with lower affinity of the order of 10^8 liters/mole necessitate the use of more than 10^{12} molecules under conventional practice. In fact all immunoassay kits marketed commercially at the present time conform to these concepts and use an amount of binding site approximating to or, more frequently, considerably in excess of V/K; indeed in certain types of kit relying on the use of labelled antibodies it is conventional to use as much binding agent as possible, binding proportions of analyte greatly exceeding 50%.

Because of the binding of substantial proportions, for example 50%, of the analyte in the liquid samples under test in such systems, the fractional occupancy of the binding sites of the binding agent is not independent of the volume of the fluid sample so that for accurate quantitative assays it is necessary to control accurately the volume of the sample, keeping it constant in all tests, whether of the sample of unknown concentration or of the standard samples of known concentration used to generate the dose response curve. Furthermore, such systems also require careful control of the amount of binding agent present in the standard and control incubation tubes. These limitations of present techniques are universally recognised and accepted.

UK Patent Application 2,099,578A discloses a device for immunoassays comprising a porous solid support to which antigens, or less frequently immunoglobulins, are bound at a plurality of spaced apart locations, said device permitting a large number of qualitative or quantitative immunoassays to be performed on the same support, for example to establish an antibody profile of a sample of human blood serum. However, although the individual locations may be in the form of so-called microdots produced by supplying droplets of antigen-containing solutions or suspensions, the number of moles of antigen present at each location is apparently still envisaged as being enough to bind essentially all of the analyte (e.g. antibody) whose concentration is to be measured that is present in the liquid sample under test. This is apparent from the fact that the quantitative method used in that application (page 3, lines 21–28) involves calibration with known amounts of immunoglobulin being applied to the support; but this means that, in the samples being tested, essentially every molecule must be extracted from the sample in order for a true comparison to be made and hence that large amounts of antigen (i.e. the binding agent in this situation) are required in each microdot, greatly in excess of the total amount of analyte (i.e. antibody in this situation) present in the sample.

SUMMARY OF THE INVENTION

The present invention involves the realisation that the use of high quantities of binding agent is neither necessary for good sensitivity in immunoassays nor is it generally desirable. If, instead of being kept as large as possible, the amount of binding agent is reduced so that only an insignificant proportion of the analyte is reversibly bound to it, generally less than 10%, usually less than 5% and for optimum results only 1 or 2% or less, not only is it no longer necessary to use an accurately controlled, constant volume for all the liquid samples (standard solutions and unknown samples) in a given assay, but it is also possible to obtain reliable and sometimes even improved estimates of analyte concentration using much less than V/K moles of binding agent binding sites, say not more than 0.1 V/K and preferably less than 0.01 V/K. For a binding agent having an equilibrium constant (K) for the analyte of the order of 10^{11} liters/mole and samples or approximately 1 ml size this is approximately equivalent to not more than 10^8, preferably less than 10^7, molecules of binding agent at each location in an individual array. If the value of K is 10^{13} liters/mole the figures are 10^6 and 10^5 molecules respectively, and if K is of the order of 10^8 liters/mole they are 10^{11} and 10^{10} molecules respectively. Below 10^2 molecules of binding agent at a single location the accuracy of the measurement would become progressively less as the fractional occupancy of the binding agent sites by the analyte would be able to change only in discrete steps as individual sites become occupied or unoccupied, but in principle at least the use of as low as 10 molecules would be permissible if an estimate with an accuracy of 10% is acceptable. Practical considerations may give rise to a preference for more than 10^4 molecules.

It will be appreciated that the abovementioned GB patent application 2,099,578A, which for quantitative estimation relies on large amounts of binding agent and essentially total sequestration of all analyte, fails to recognise the advance achieved by the present invention, which instead relies on a different analytical principle requiring measurement of the fractional occupancy of the binding agent and which thus requires only a very low proportion of the total analyte molecules present to be sequestered from the sample.

Following the recognition that the use of such small amounts of binding agent is permissible, it becomes feasible to place the binding agent required for a single concentration measurement on a very small area of a solid support and hence to place in juxtaposition to one another but at spatially separate points on a single solid support a wide variety of different binding agents specific for different analytes which are or may be present simultaneously in a liquid to be analysed. Simultaneous exposure of each of the separate points to the liquid to be analysed will cause each binding agent spot to take up the analyte for which it is specific to an extent (i.e. fractional binding site occupancy) representative of the analyte concentration in the liquid, provided only that the volume of solution and the analyte concentration therein are large enough that only an insignificant fraction (generally less than 10%, usually less than 5%) of the analyte is bound to the point. The fractional binding site occupancy for each binding agent can then be determined using separate site-recognition reagents which recognise either the unfilled binding sites or filled binding sites of the different binding agents and which are labelled with markers enabling the concentration levels of the separate reagents bound to the different binding agents to be measured, for example fluorescent markers. Such measurements may be performed consecutively, for example using a laser which scans across the support, or simultaneously, for example using a photographic plate, depending on the nature of the labels. Other imaging devices such as a television camera can also be used where appropriate Because the binding agents are spatially separate from one another it is possible to use only a small number of different marker labels or even the same marker label throughout and to scan each binding agent location separately to determine the presence and concentration of the label. By use of the invention considerably more than 3 analyses can be performed with a single exposure of the solid support with liquid to be analysed, for example 10, 20, 30, 50 or even up to 100 or several hundreds of analyses.

Overall, therefore, the present invention provides a method for determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising:

loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid and is specific for that analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart locations such that each location has not more than 0.1 V/K moles of a single binding agent, where K liters/mole is the equilibrium constant of the binding agent for the analyte,

contacting the loaded support means with the liquid sample to be analysed such that each of the spaced apart locations is contacted in the same operation with the liquid sample, the amount of liquid used in the sample being such that only an insignificant proportion of any analyte present in the liquid sample becomes bound to the binding agent specific for it, and

measuring a parameter representative of the fractional occupancy by the analytes of the binding agents at the spaced apart locations by a competitive or non-competitive assay technique using a site-recognition reagent for each binding agent capable of recognising either the unfilled binding sites or the filled binding sites on the binding agent, said site-recognition reagent being labelled with a marker enabling the amount of said reagent in the particular location to be measured.

The invention also provides a device for use in determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising a solid support means having located thereon at a plurality of spaced apart locations a plurality of different binding agents, each binding agent being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for that analyte as compared to the other components of the liquid sample, each location having not more than 0.1 V/K, preferably less than 0.01 V/K, moles of a single binding agent, where K liters/mole is the equilibrium constant of that binding agent for reaction with the analyte to which it is specific.

A kit for use in the method according to the invention comprises a device according to the invention, a plurality of standard samples containing known concentrations of the analytes whose concentrations in the liquid sample are to be measured and a set of labelled site-recognition reagents for reaction with filled or unfilled binding sites on the binding agents.

In arriving at the method of the invention, I have found that, generally speaking, for antibodies having an affinity constant K liters/mole for an antigen, the relationship between the antibody concentration and the fractional occupancy of the binding sites at any particular antigen concentration and the relationship between the antibody concentration and the percentage of antigen bound to the binding sites at any particular antigen concentration follow the same curves provided that the antibody concentrations and the antigen concentrations are each expressed in terms of fractions or multiples of 1/K.

BRIEF DESCRIPTION OF THE DRAWING

The principle underlying the method of the invention may be better understood by reference to the accompanying drawing which is a graph representing two sets of curves plotting the relationship between antibody concentration and the fractional occupancy of the binding sites at certain prescribed antigen concentrations and the relationship between antibody concentration and the percentage of antigen bound to the binding sites at the same prescribed antigen concentrations. Each curve relates to the antibody concentration [Ab], expressed in terms of 1/K, plotted along the x-axis. For the set of curves which remain constant or decline with increasing [Ab], the y-axis represents the fractional occupancy (F) of binding sites on the antibody by the antigen; for the second set, the y-axis represents the percentage (be) of antigen bound to those binding sites. The individual curves in each set represent the relationships corresponding to four different antigen concentrations [An] expressed in terms of K, namely 10/K, 1.0/K, 0.1/K and 0.01/K. The curves show that as [Ab] falls F reaches an essentially constant level, the value of which is dependent on [An].

DETAILED DESCRIPTION

The choice of a solid support is a matter to be left to the user. Preferably the support is non-porous so that the binding agent is disposed on its surface, for example as a monolayer. Use of a porous support may cause the binding agent, depending on its molecular size, to be carried down into the pores of the support where its exposure to the analyte whose concentration is to be determined may likewise be affected by the geometry of the pores, so that a false reading may be obtained. Porous supports such as nitrocellulose paper dotted with spots of binding agent are therefore less preferred. Unlike the supports used in GB 2,099,578A, which seem to need to be porous because of the large number of molecules to be attached, the supports for use in the present invention use much smaller quantities and therefore need not be porous. The non-porous supports may, for example be of plastics material or glass, and any convenient rigid plastics material may be used. Polystyrene is a preferred plastics material, although other polyolefins or acrylic or vinyl polymers could likewise be used.

The support means may comprise microbeads, e.g. of such a plastics material, which can be coated with uniform layers of binding agent and retained in specified locations, e.g. hollows, on a support plate. Alternatively the material may be in the form of a sheet or plate which is spotted with an array of dots of binding agent. It can be advantageous for the configuration of the support means to be such that liquid samples of approximately the volume V liters are readily retained in contact with the plurality of spaced apart locations marked with the different binding agents. For example, the spaced apart locations may be arranged in a well in the support means, and a plurality of wells, each provided with the same group of different binding agents in spaced apart locations, can be linked together to form a microtitre plate for use with a plurality of samples.

When the support means is to be used in conjunction with a measuring system involving light scanning, the material, e.g. plastics, for the support is desirably opaque to light, for example it may be filled with an opacifying material which may inter alia be white or black, such as carbon black, when the signals to be measured from the binding agent or the site-recognition reagent are light signals, as from fluorescent or luminescent markers. In general, reflective materials are preferred in this case to enhance light collection in the detecting instrument or photographic plate. The final choice of optimum material is governed by its ability to attach the binding agent to its surface, its absence of background signal emission and its possession of other properties tending to maximise the signal/noise ratio for the particular marker or markers attached to the binding agent situated on its surface. Very satisfactory results have been obtained in the Examples described below by the use of a white opaque polystyrene microtitre plate commercially available from Dynatech under the trade name White Microfluor microtitre wells.

The binding agents used may be binding agents of different specificity, that is to say agents which are specific to different analytes, or two or more of them may be binding agents of the same specificity but of different affinity, that is to say agents which are specific to the same analyte but have different equilibrium constants K for reaction with It. The latter alternative is particularly useful where the concentration of analyte to be assayed in the unknown sample can vary over considerable ranges, for example 2 or 3 orders of magnitude, as in the case of HCG measurement in urine of pregnant women, where it can vary from 0.1 to 100 or more IU/ml.

The binding agents used will preferably be antibodies, more preferably monoclonal antibodies. Monoclonal antibodies to a wide variety of ingredients of biological fluids are commercially available or may be made by known techniques. The antibodies used may display conventional affinity constants, for example from 10^8 or 10^9 liters/mole upwards, e.g. of the order of 10^{10} or 10^{11} liters/mole, but high affinity antibodies with affinity constants of 10^{12}–10^{13} liters/mole can also be used. The invention can be used with such binding agents which are not themselves labelled. However, it is also possible and frequently desirable to use labelled binding agents so that the system binding agent/analyte/site-recognition reagent includes two different labels of the same type, e.g. fluorescent, chemiluminescent, enzyme or radioisotopic, one on the binding agent and one on the site-recognition reagent. The measuring operation then measures the ratio of the intensity of the two signals and thus eliminates the need to place the same amount of labelled binding agent on the support when measuring signals from standard samples for calibration purposes as when measuring signals from the unknown samples. Because the system depends solely on measurement of a ratio representative of binding site occupancy, there is also no need to measure the signal from the entire spot but scanning only a portion is sufficient. Each binding agent is preferably labelled with the same label but different labels can be used.

The binding agents may be applied to the support in any of the ways known or conventionally used for coating binding agents onto supports such as tubes, for example by contacting each spaced apart location on the support with a solution of the binding agent in the form of a small drop, e.g. 0.5 microliter, on a 1 mm^2 spot, and allowing them to remain in contact for a period of time before washing the drops away. A roughly constant small fraction of the binding agent present in the drop becomes adsorbed onto the support as a result of this procedure. It is to be noted that the coating density of binding agent on the microspot does not need to be less than the coating density in conventional antibody-coated tubes; the reduction in the number of molecules on each spot may be achieved solely by reduction of the size of the spot rather than the coating density. A high coating density is generally desirable to maximise signal/noise ratios. The sizes of the spots are advantageously less than 10 mm^2, preferably less than 1 mm^2. The separation is desirably, but not necessarily, 2 or 3 times the radius of the spot, or more. These suggested geometries can nevertheless be changed as required, being subject solely to the limitations on the number of binding agent molecules in each spot, the minimum volume of the sample to which the array of spots will be exposed and the means locally available for conveniently preparing an array of spots in the manner described.

Once the binding agents have been coated onto the support it is conventional practice to wash the support, in the case of antibodies as binding agents, with a solution containing albumen or other protein to saturate all remaining non-specific adsorption sites on the support and elsewhere. To confirm that the amount of binding agent in an individual spot will be less than the maximum amount (0.1 V/K) required to conform to the principle of the present invention, the amount of binding agent present on any individual site can be checked by labelling the binding agent with a detectable marker of known specific activity (i.e. known amount of marker per unit weight of binding agent) and measuring the amount of marker present. Thus, if the use of labelled binder is not desired on the solid support used in the method of the invention the binding agent can nevertheless be labelled in a trial experiment and identical conditions to those found in that trial to give rise to correct loadings of binding agent can be used to apply unlabelled binding agent to the supports to be actually used.

The minimum size of the liquid sample (V liters) is correlated with the number of mole of binding agent (less than 0.1 V/K) so that only an insignificant proportion of the analyte present in the liquid sample becomes bound to the binding agent. This proportion is as a general rule less than 10%, usually less than 5% and desirably 1 or 2% or less, depending on the accuracy desired for the assay (greater accuracy being obtained, other things being equal, when smaller proportions of analyte are bound) and the magnitude of other error-introducing factors present. Sample sizes of the order of one or a few ml or less, e.g. down to 100 microliters or less, are often preferred, but circumstances may arise when larger volumes are more conveniently assayed, and the geometry may be adjusted accordingly. The sample may be used at its natural concentration level or if desired it may be diluted to a known extent.

The site-recognition reagents used in the method according to the invention may themselves be antibodies, e.g. monoclonal antibodies, and may be anti-idiotypic or anti-analyte antibodies, the latter recognising occupied sites. Alternatively, for example for analytes of small molecular size such as thyroxine (T4), unoccupied sites may be recognised using either the analyte itself, appropriately labelled, or the analyte covalently coupled to another molecule—e.g. a protein molecule—which is directly or indirectly labelled. The site-recognition reagents may be labelled directly or indirectly with conventional fluorescent labels such as fluorescein, rhodamine or Texas Red or materials usable in time-resolved pulsed fluorescence such as europium and other lanthanide chelates, in a conventional manner. Other labels such as chemiluminescent, enzyme or radioisotopic labels may be used if appropriate. Each site-recognition reagent is preferably labelled with the same label but different labels can be used in different reagents. The site-recognition reagents may be specific for a single one of the binding agent/analyte spots in each group of spots or in certain circumstances, as with glycoprotein hormones such as HCG and FSH which have a common binding site, they may be cross-reacting reagents able to react with occupied binding sites in more than one of the spots.

In the assay technique the signals representative of the fractional occupancy of the binding agent in the test samples of unknown concentrations of the analytes can be calibrated by reference to dose response curves obtained from standard samples containing known concentrations of the same analytes. Such standard samples need not contain all the analytes together, provided that each of the analytes is present in some of the standard samples. Fractional occupancy may be measured by estimating occupied binding sites (as with an anti-analyte antibody) or unoccupied binding sites (as with an anti-idiotypic antibody), as one is the converse of the other. For greater accuracy it is desirable to measure the fraction which is closer to zero because a change in fractional occupancy of 0.01 is proportionately greater in this case, although for fractional occupancies in the range 25–75% either alternative is generally satisfactory.

In that embodiment of the present invention which relies on two fluorescent markers, the measurement of relative intensity of the signals from the two markers, one on the binding agent and the other on the site recognition reagent, may be carried out by a laser scanning confocal microscope such as a Bio-Rad Lasersharp MRC 500, available from Bio-Rad Laboratories Ltd., and having a dual channel detection system. This instrument relies on a laser beam to scan the dots or the like on the support to cause fluorescence of the markers and wavelength filters to distinguish and measure the amounts of fluorescence emitted. Time-resolved fluorescence methods may also be used. Interference (so-called crosstalk) between the two channels can be compensated for by standard corrections if it occurs or conventional efforts can be made to reduce it. Discrimination of the two fluorescent signals emitted by the dual-labelled spots is accomplished in the present form of this instrument, by filters capable of distinguishing the characteristic wavelength of the two fluorescent emissions; however, fluorescent substances may be distinguished by other physical characteristics such as differing fluorescence decay times, bleaching times, etc., and any of these means may be used, either alone or in combination, to differentiate between two fluorophores and hence permit measurement of the ratio of two fluorescent labelled entities (binding agent and site-recognition reagent) present on an individual spot, using techniques well known in the fluorescence measurement field. When only one fluorescent label is present the same techniques may be used, provided that care is taken to scan the entire spot in each case and the spots contain essentially the same amount of binding agent from one assay to the next when the unknown and standard samples are used.

In the case of other labels, such as radioisotopic labels, chemiluminescent labels or enzyme labels, analogous means of distinguishing the individual signals from one or from each of a pair of such labels are also well known. For example two radioisotopes such as ^{125}I and ^{131}I may be readily distinguished on the basis of the differing energies of their respective radioactive emissions. Likewise it is possible to identify the products of two enzyme reactions, deriving from dual enzyme-labelled antibody couplets, these being e.g. of different colours, or two chemiluminescent reactions, e.g. of different chemiluminescent lifetime or wavelength of light emission; by techniques well known in the respective fields.

The invention may be used for the assaying of analytes present in biological fluids, for example human body fluids such as blood, serum, saliva or urine. They may be used for the assaying of a wide variety of hormones, proteins, enzymes or other analytes which are either present naturally in the liquid sample or may be present artificially such as drugs, poisons or the like.

For example, the invention may be used to provide a device for quantitatively assaying a variety of hormones relating to pregnancy and reproduction, such as FSH, LH, HCG, prolactin and steroid hormones (e.g. progesterone, estradiol, testosterone and androstene-dione), or hormones of the adrenal pituitary axis, such as cortisol, ACTH and aldosterone, or thyroid-related hormones, such as T4, T3, and TSH and their binding protein TBG, or viruses such as hepatitis, AIDS or herpes virus, or bacteria, such as staphylococci, streptococci, pneumococci, gonococci and enterococci, or tumour-related peptides such as AFP or CEA, or drugs such as those banned as illicit improvers of athletes' performance, or food contaminants. In each case the binding agents used will be specific for the analytes to be assayed (as compared with others in the sample) and may be monoclonal antibodies therefor.

Further details on the methodology are to be found in my International Patent Publication WO88/01058, the contents of which are incorporated herein by reference.

The invention is illustrated by the following Examples.

EXAMPLE 1

An anti-TNF (tumour necrosis factor) antibody having an affinity constant for TNF at 25° C. of about 1×10^{9} liters/mole is labelled with Texas Red. A solution of the antibody at a concentration of 80 micrograms/ml is formed and 0.5 microliter aliquots of this solution are added in the form of droplets one to each well of a Dynatech Microfluor (opaque white) filled polystyrene microtitre plate having 12 wells.

An anti-HCG (human chorionic gonadotropin) antibody having an affinity constant for HCG at 25° C. of about 6×10^{8} liters/mole is also labelled with Texas Red. A solution of the antibody at a concentration of 80 micrograms/ml is formed and 0.5 microliter aliquots of this solution are added in the form of droplets one to each well of the same Dynatech Microfluor microtitre plate.

After addition of the droplets the plate is left for a few hours in a humid atmosphere to prevent evaporation of the droplets. During this time some of the antibody molecules in the droplets become adsorbed onto the plate. Next, the wells are washed several times with a phosphate buffer and then they are filled with about 400 microliters of a 1% albumen solution and left for several hours to saturate the residual binding sites in the wells. Thereafter they are washed again with phosphate buffer.

The resulting plate has in each of its wells two spots each of area approximately 1 mm^2. Measurement of the amount of fluorescence shows that in each well one spot contains about 5×10^{9} molecules of anti-TNF antibody and the other contains about 5×10^{9} molecules of anti-HCG antibody. The wells are designed for use with liquid samples of volume 400 microliters, so that 0.1 V/K is 4×10^{-14} moles (equivalent to 2.4×10^{10} molecules) for the anti-TNF antibody and 7×10^{-14} moles (equivalent to 4×10^{10} molecules) for the anti-HCG antibody.

EXAMPLE 2

A microtitre plate prepared as described in Example 1 is used in an assay for an artificially produced solution containing TNF and HCG. A test sample of the solution, amounting to about 400 microliters, is added to one of the wells and allowed to incubate for several hours. About 400 microliters of various standard solutions containing known concentrations (0.02, 0.2, 2 and 20 ng/ml) of TNF or HCG are added to other wells of the plate and also allowed to incubate for several hours. The wells are then washed several times with buffer solution.

As site-recognition reagents there are used for the TNF spots an anti-TNF antibody having an affinity constant for TNF at 25° C. of about 1×10^{10}liters/mole and for the HCG spots an anti-HCG antibody having an affinity constant for HCG at 25° C. of about 1×10^{11} liters/mole. Both antibodies are labelled with fluorescein (FITC). 400 microliter aliquots of solutions of these labelled antibodies are added to the wells and allowed to stand for a few hours. The wells are then washed with buffer.

The resulting fluorescence ratio of each spot is quantified with a Bio-Rad Lasersharp MRC 500 confocal microscope. From the standard solutions dose response curves for TNF and HCG are built up, the figures for TNF being as follows:

TNF concentration ng/ml	$\frac{\text{FITC fluorescence}}{\text{Texas Red fluorescence}}$ on TNF spot
0.02	1.1
0.2	4.6
2	7.9
20	42.5

and those for HCG being as follows:

HCG concentration ng/ml	$\frac{\text{FITC fluorescence}}{\text{Texas Red fluorescence}}$ on HCG spot
0.02	1.8
0.2	7.2
2	16.0
20	28.2

The artificially produced solution was found to give ratio readings of 5.9 on the TNF spot and 10.5 on the HCG spot, correlating well with the actual concentrations of TNF (0.5 ng/ml) and HCG (0.5 ng/ml) obtained from the dose response curves.

EXAMPLE 3

Using similar procedures to those outlined in Example 1 a microtitre plate containing spots of labelled anti-T4 (thyroxine) antibody (affinity constant about 1×10^{11} liters/mole at 25° C.), labelled anti-TSH (thyroid stimulating hormone) antibody (affinity constant about 5×10^{9} liters/mole at 25° C.) and labelled anti-T3 (triiodothyronine) antibody (affinity constant about 1×10^{11} liters/mole at 25° C.) in each of the individual wells is produced, the spots containing less than 1×10^{-12} V moles of anti-T4 antibody or less than 2×10^{-11} V moles of anti-TSH antibody or less than 1×10^{-12} V moles of anti-T3 antibody.

The developing antibody (site-recognition reagent) for the TSH assay is an anti-TSH antibody with an affinity constant for TSH of 2×10^{10} liters/mole at 25° C. This antibody is labelled with fluorescein (FITC). The site-recognition reagents for the T4 and T3 assays are T4 and T3 coupled to poly-lysine and labelled with FITC, and they recognise the unfilled sites on their respective first antibodies.

Using 400 microliter aliquots of standard solutions containing various known amounts of T4, T3 and TSH, dose response curves are obtained by methods analogous to those in Example 2, correlating fluorescence ratios with T4, T3 and TSH concentrations. The plate is used to measure T4, T3 and TSH levels in serum from human patients with good correlation with the results obtained by other methods.

EXAMPLE 4

Using similar procedures to those outlined in Example 1 a microtitre plate containing spots of first labelled anti-HCG antibody (affinity constant about 6x10^8 liters/mole at 25° C.), second labelled anti-HCG antibody (affinity constant about 1.3x10^{11} liters/mole at 25° C.) and labelled anti-FSH (follicle stimulating hormone) antibody (affinity constant about 1.3x10^8 liters/mole at 25° C.) in each of the individual wells is produced, the spots each containing less than 0.1 V/K moles of the respective antibody. A cross-reacting (alpha subunit) monoclonal antibody 8D10 with an affinity constant of 1x10^{11} liters/mole is used as a common developing antibody for both the HCG and the FSH assays.

Using 400 microliter aliquots of standard solutions containing various known concentrations of HCG and FSH, dose response curves are obtained by methods analogous to those in Example 2, correlating fluorescence ratios with HCG and FSH concentrations, the curve obtained with the higher affinity anti-HCG antibody giving more concentration-sensitive results at the lower HCG concentrations whereas the curve from the lower affinity anti-HCG antibody is more concentration-sensitive at the higher HCG concentrations. The plate is used to measure HCG and FSH concentrations in the urine of women in pregnancy testing, giving good correlations with results obtained by other means and achieving effective concentration measurements for HCG over a concentration range of two or three orders of magnitude by correct choice of the best HCG spot and dose response curve.

Production of Labelled Antibodies

The labelling of the antibodies with fluorescent labels can be carried out by a well known and standard technique, see Leslie Hudson and Frank C. Hay, "Practical Immunology", Blackwell Scientific Publications (1980), pages 11–13, for example as follows:

The monoclonal antibody anti-FSH 3G3, an FSH specific (beta subunit) antibody having an affinity constant (K) of 1.3x10^8 liters per mole, was produced in the Middlesex Hospital Medical School, and was labelled with TRITC (rhodamine isothiocyanate) or Texas Red, giving a red fluorescence.

The monoclonal antibody anti-FSH 8D10, a cross-reacting (alpha subunit) antibody having an affinity constant (K) of 1x10^{11} liters per mole, was likewise produced in the Middlesex Hospital Medical School and was labelled with FITC (fluorescein isothiocyanate), giving a yellow-green fluorescence.

The general procedure used involved ascites fluid purification (ammonium sulphate precipitation and T-gel chromatography) followed by labelling, according to the following steps:

1.a. Ammonium sulphate purification

1. Add 4.1 ml saturated ammonium sulphate solution to 5 ml antibody preparation (culture supernatant or 1:5 diluted ascites fluid) under constant stirring (45% saturation).
2. Continue stirring for 30–90 min. Centrifuge at 2500 rpm for 30 min.
3. Discard the supernatant and dissolve the precipitate in PBS (final volume 5 ml.). Repeat Steps 1 and 2, OR
4. Add 3.6 ml saturated ammonium sulphate (40% saturation) under constant stirring. Repeat Step 2.
5. Discard the supernatant and dissolve the pellet in the desired buffer.
6. Dialyse overnight in cold against the same buffer (using fresh, boiled-in-d/w dialysis bag).
7. Determine the protein concentration either at A_{280} or by Lowry estimation.

1.b. T-gel Chromatography: (Buffer: 1M Tris-Cl, pH 7.6. Solid potassium sulphate)

1. Clear 2 ml of ascites fluid by centrifugation at 4000 rpm.
2. Add 1M Tris-Cl solution to achieve final concentration of 0.1M.
3. Add sufficient amount of solid potassium sulphate. Final concentration=0.5M.
4. Apply the ascite fluid to the T-gel column.
5. Wash the column with 0.1M Tris-Cl buffer containing 0.5M potassium sulphate, until protein profile (at A_{280}) returns to zero.
6. Elute the absorbed protein using 0.1M Tris-Cl buffer as the eluant.
7. Pool the fractions containing antibody activity and concentrate using Amicon 30 concentrater.
8. If HPHT purification is to be carried out, use HPHT chromatography Starting buffer during Step 7.

2. Labelling of Antibodies FITC/TRITC conjugation

1. Dialyse the purified 1 g protein into 0.25M Carbonate-bicarbonate buffer, pH 9.0 to a concentration of 20 mg/ml.
2. Add FITC/TRITC to achieve a 1:20 ratio with protein (i.e. 0.05 mg for every 1 mg of protein).
3. Mix and incubate at 4° C. for 16–18 hrs.
4. Separate the conjugated protein from unconjugated by:
   a. Sephadex G-25 chromatography for FITC label,

or

   b. DEAE-Sephacel chromatography for TRITC/FITC label.

Buffer system:

PBS for (a).

0.005M Phosphate, pH 8.0 and 0.18M

Phosphate, pH 8.0 for (b).

Calculation of *FITC*: Protein coupling ratio: –

$$\frac{2.87 \times O.D.\ 495\ \text{nm}}{O.D.\ 280\ \text{nm} - 0.35 \times O.D.\ 495\ \text{nm}}$$

EXAMPLE 4

Regents

1 TSH standards from the National Institute for Biological Standards and Control

2 TSH-free Serum for making up TSH standards

3 ^{125}I-labelled TSH

4 Anti-TSH monoclonal antibodies from The Scottish Antibody Production Unit

5 Phosphate buffer, 0.1M, pH 7.4

6 Tris-HCl buffer, 0.05M, pH 7.6, containing 0.5% bovine serum albumin (BSA), 0.05% Tween 20 and 0.1% sodium azide

7 Wash buffer: Phosphate buffer, 0.1M, pH 7.4, containing 0.1% Tween 20 and 0.1% sodium azide

8 Black microtitre strips from Dynatech

9 SuperBlock from Pierce

A. Protocol and Conditions for the Radioimmunoassay of Thyroid Stimulating Hormone (TSH)

1. An aliquot of 50 μl of 50 μg/ml anti-TSH monoclonal antibody in phosphate buffer was added to microtitre wells and incubated for 1 hour at room temperature.
2. The microtitre wells were washed with phosphate buffer, blocked with SuperBlock for 30 minutes at room temperature and then washed again.
3. An aliquot of 100 μl of TSH standards made up in TSH-free serum (to yield final concentrations of 0, $1x^{-9}$, $2x10^{-9}$, $4x10^{-9}$, $8x10^{-9}$, $12x10^{-9}$, $16x10^{-9}$ and 20x10 M/L) or unknown serum samples and 100 μl of ^{125}I-labelled TSH in Tris-HCl assay buffer were added to triplicate anti-TSH monoclonal antibody coated microtitre wells, shaken for 1 hour at room temperature, washed with wash buffer and counted for radioactivity. The concentration of TSH in the unknown samples can be read from the standard curve.

The incubation period of 1 hour for the assay is far less than the time required for the binding reaction to go to equilibrium, but, provided the standards are measured under the same conditions, the unknown sample can be measured against those standards. The effective affinity constant for the antibody will of course be that which pertains after 1 hour incubation and under the same conditions as the assay itself.

B. Procedure for Obtaining the Affinity Constant K of the Anti-TSH Monoclonal Antibody Used in a Radioimmunoassay Performed Under the Conditions Described in (A)

1. An aliquot of 50 Al of 50 μg/ml anti-TSH monoclonal antibody in phosphate buffer was added to microtitre wells and incubated for 1 hour at room temperature.
2. The microtitre wells were washed with phosphate buffer, blocked with SuperBlock for 30 minutes at room temperature and then washed again.
3. An aliquot of 100 μl of TSH standards made up in TSH-free serum (to yield final concentrations of 0, $1x10^{-9}$, $2x10^{-9}$, $4x10^{-9}$, $8x10^{-9}$, $12x10^{-9}$, $16x10^{-9}$ and $20x10^{-9}$ M/L) and 100 μl of ^{125}I-labelled TSH in Tris-HCl assay buffer were added to triplicate antibody coated microtitre wells, shaken for 1 hour at room temperature, washed with wash buffer and counted for radioactivity.
4. A standard Scatchard plot of Bound/Free vs. Bound TSH was used to obtain the affinity constant K for the monoclonal anti-TSH antibody.

C. A TSH Assay Using an Amount of Capture Antibody ≦0.1 V/K and Deposited on the Solid-Phase as Microspots

Since the assay volume V is 0.2 ml or $2x10^{-4}$ L and the affinity constant K of the anti-TSH capture antibody used under conditions described in (B) was found to be $1.1x10^{8}$ L/M, therefore the maximum amount of capture antibody allowed in the assay under ambient analyte condition

$$
\begin{aligned}
&-0.1\ V/K\\
&= (0.1 \times 2 \times 10^{-4})/1.1 \times 10^{8} M\\
&= 1.8 \times 10^{-13} M
\end{aligned}
$$

Or a capture antibody concentration of $9x10^{10}$ M/L.

Assay Protocol:

1. A 0.5 μl droplet of a monoclonal anti-TSH capture antibody in phosphate buffer and at a concentration of 200 μg/ml was added to each microtitre well and aspirated instantly. This procedure resulted in antibody microspots with a coated area of approximately 10^{6} μm^{2}.

$$
\begin{aligned}
&\text{Molar amount of coated antibody on microspot}\\
&= (\text{coated area} \times \text{antibody density})/\text{Avogadro Number} -\\
&\quad (10^{4} \times 10^{6})/(6.01 \times 10^{23}) M\\
&= 1.7 \times 10^{-14} M
\end{aligned}
$$

or a capture antibody concentration of $0.85x10^{-10}$ M/L.

2. The microtitre wells were washed with phosphate buffer and the unreacted sites blocked with SuperBlock for 30 minutes at room temperature and then washed again with phosphate buffer.
3. 100 μl of TSH standards (made up in TSH-free serum) or unknown samples plus 100 μl of Tris-HCl assay buffer were added to triplicate microtitre wells, shaken for 1 hour at room temperature and washed with wash buffer.
4. The TSH bound sites were back-titrated using fluorescent labelled anti-TSH developing monoclonal antibody raised against a different site on the TSH molecule and complementary to the capture antibody deposited as microspot on the solid-phase. An aliquot of 200 μl of the developing antibody in Tris-HCl assay buffer was added to the microtitre wells, shaken for 1 hour at room temperature, washed with wash buffer, scanned with a BioRad laser scanning confocal microscope and the amount of fluorescence on the microspots and the amount of fluorescence on the microspots quantified. The concentration of TSH in the unknown samples were read from the standard curve.

Although, for the purpose of illustration, the affinity constant of the antibody was measured under the assay conditions, in practice, in many cases it may not be necessary actually to perform such a measurement, so long as it is obvious, having regard to the details of the assay in question, that the amount of capture antibody used on any spot is going to be less than 0.1 V/K.

What is claimed is:

1. A method for determining the ambient concentration of an analyte of interest among a plurality of analytes in a liquid sample of volume V liters, comprising:

loading a plurality of different binding agents, each being labelled with a marker and being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for said analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart small spots such that not more than 0.1 V/K moles of binding agent are present on any spot, where K liters/mole is the affinity constant of said binding agent for said analyte;

contacting the loaded support means with the liquid sample to be analyzed, such that each of the spots is contacted in the same step with said liquid sample, the amount of liquid used in said sample being such that only an insignificant proportion of any analyte present in said liquid sample becomes bound to said binding agent specific for said analyte; .

contacting the support with a site-recognition reagent specific for each binding agent in a competitive or non-competitive technique, the site-recognition reagent being capable of recognizing either the unfilled binding sites or the filled binding sites on said binding agent, said site-recognition reagent being labelled with a marker different from the marker on said binding agent, and measuring a ratio of signals from said markers on the site recognition reagent and the binding reagent from at least a part of the spot, from which the analyte to interest is determined.

2. A method according to claim 1, wherein the markers on the site-recognition reagent and the binding reagent are fluorescent markers.

3. A method according to claim 2, wherein the ratio of signals is measured using a laser scanning confocal microscope.

4. A method for determining the fractional binding site occupancy of a plurality of binding agents by a plurality of analytes in a liquid sample of V liters, comprising:

(a) loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for said analyte as compared to the other components of the liquid sample, onto a support at a plurality of spaced apart small spots such that each spot has a high coating density of one of said binding agents but not more than 0.1 V/K moles of binding agent are present on any one spot, where K liters/mole is the affinity constant of said binding agent for said analyte;

(b) contacting the loaded support with the liquid sample to be analyzed, such that each of the spots is contacted in the same step with said liquid sample, the amount of liquid used in said sample being such that only an insignificant proportion of any analyte present in said liquid sample becomes bound to said binding agent specific for said analyte; and

(c) thereafter contacting the loaded support with site-recognition reagents which recognize either the unfilled binding sites or filled binding sites of that binding agent, the site-recognition reagents being labelled with markers from which the fractional binding site occupancy for each binding agent is determined.

5. The method of claim 4, wherein the site-recognition reagents are labelled with fluorescent markers.

6. The method of claim 4, wherein the presence of the site-recognition reagents on each respective binding agent is determined consecutively.

7. The method of claim 4, wherein the presence of the site-recognition reagents on each respective binding agent is determined simultaneously.

8. The method of claim 4, further comprising, after step (c), calculating the concentration level of each reagent using the determined value of the fractional binding site occupancy.

9. A method for detecting a plurality of analytes in a liquid sample of volume V liters, comprising:

loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for said analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart small spots such that each spot has a high coating density of one of said binding agents but not more than 0.1 V/K moles of binding agent are present on any spot, where K liters/moles is the affinity constant of said binding agent for said analyte;

contacting the loaded support means with the liquid sample to be analyzed, such that each of the spots is contacted in the same step with said liquid sample, the amount of liquid used in said sample being such that only an insignificant proportion of any analyte present in said liquid sample becomes bound to said binding agent specific for said analyte;

contacting the support with a site-recognition reagent specific for each binding agent in a competitive or non-competitive technique, the site-recognition reagent being capable of recognizing either the unfilled binding sites or the filled binding sites on said binding agent, said site-recognition reagent being labelled with a marker; and

measuring the signal from the marker of the site-recognition reagent in a particular location to detect the presence of said plurality of analytes in said sample.

10. A method as claimed in claim 9, wherein each of said spots has a size of less than 1 mm^2.

11. A method as claimed in claim 10, wherein each of said spots contains more than 10^4 molecules of binding agent.

12. A method as claimed in claim 11, wherein each of said spots has less than 0.01 V/K moles of binding agent.

13. A method as claimed in claim 11, wherein said binding agents used have affinity constants for said analytes of from 10^8 to 10^{13} liters per mole.

14. A method as claimed in claim 11, wherein said binding agents used have affinity constants for said analytes of the order of $10^{10}$to 10^{11} liters per mole.

15. A method as claimed in claim 11, wherein the volume of said liquid sample is not more than 0.1 liter.

16. A method as claimed in claim 11, wherein the volume of said liquid sample is 400 to 1000 microliters.

17. A method as claimed in claim 9, wherein said binding agents loaded onto said support means are antibodies for the analytes whose concentrations are to be determined.

* * * * *

Exhibit O of J. C. Rockett Declaration
dated 12/04/03
In USSN 09/937,060

US005432099A

United States Patent [19]

Ekins

[11] **Patent Number: 5,432,099**

[45] **Date of Patent: Jul. 11, 1995**

[54] **DETERMINATION OF AMBIENT CONCENTATION OF SEVERAL ANALYTES**

[75] Inventor: **Roger P. Ekins**, London, Great Britain

[73] Assignee: **Multilyte Limited**, United Kingdom

[21] Appl. No.: **984,264**

[22] Filed: **Dec. 1, 1992**

Related U.S. Application Data

[63] Continuation of Ser. No. 460,878, filed as PCT/GB88/00649, Aug. 5, 1988.

[30] **Foreign Application Priority Data**

Feb. 10, 1988 [GB] United Kingdom 8803000

[51] **Int. Cl.**^{6} **G01N 33/543**; G01N 33/537; G01N 33/533

[52] **U.S. Cl.** **436/518**; 435/7.1; 435/7.92; 435/973; 436/501; 436/517

[58] **Field of Search** 435/7.1, 7.92, 973; 436/501, 517, 518

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,591,570	5/1986	Chang	436/518
5,096,807	3/1992	Leaback	435/6

FOREIGN PATENT DOCUMENTS

8401031	5/1984	WIPO	G01N 33/54

OTHER PUBLICATIONS

Ekins et al., "Multianalyte testing", Clin Chem. 39: 369-370 (1992).
Dudley et al., "Guidelines for Immunoassay Data Processing," Clin. Chem. 31(1264-1271) (1985).

Primary Examiner—Christine M. Nucker
Assistant Examiner—M. P. Woodward
Attorney, Agent, or Firm—Dann, Dorfman, Herrell and Skillman

[57] **ABSTRACT**

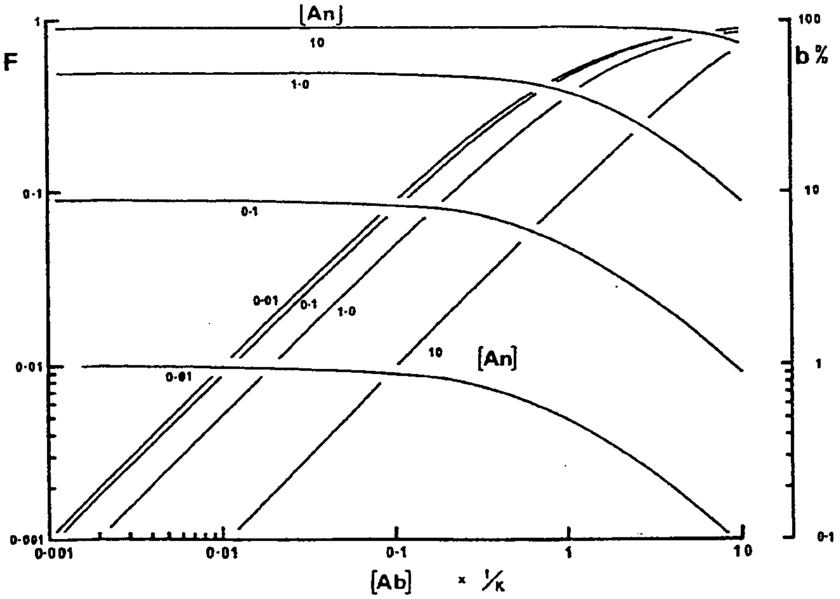
A method for determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprises

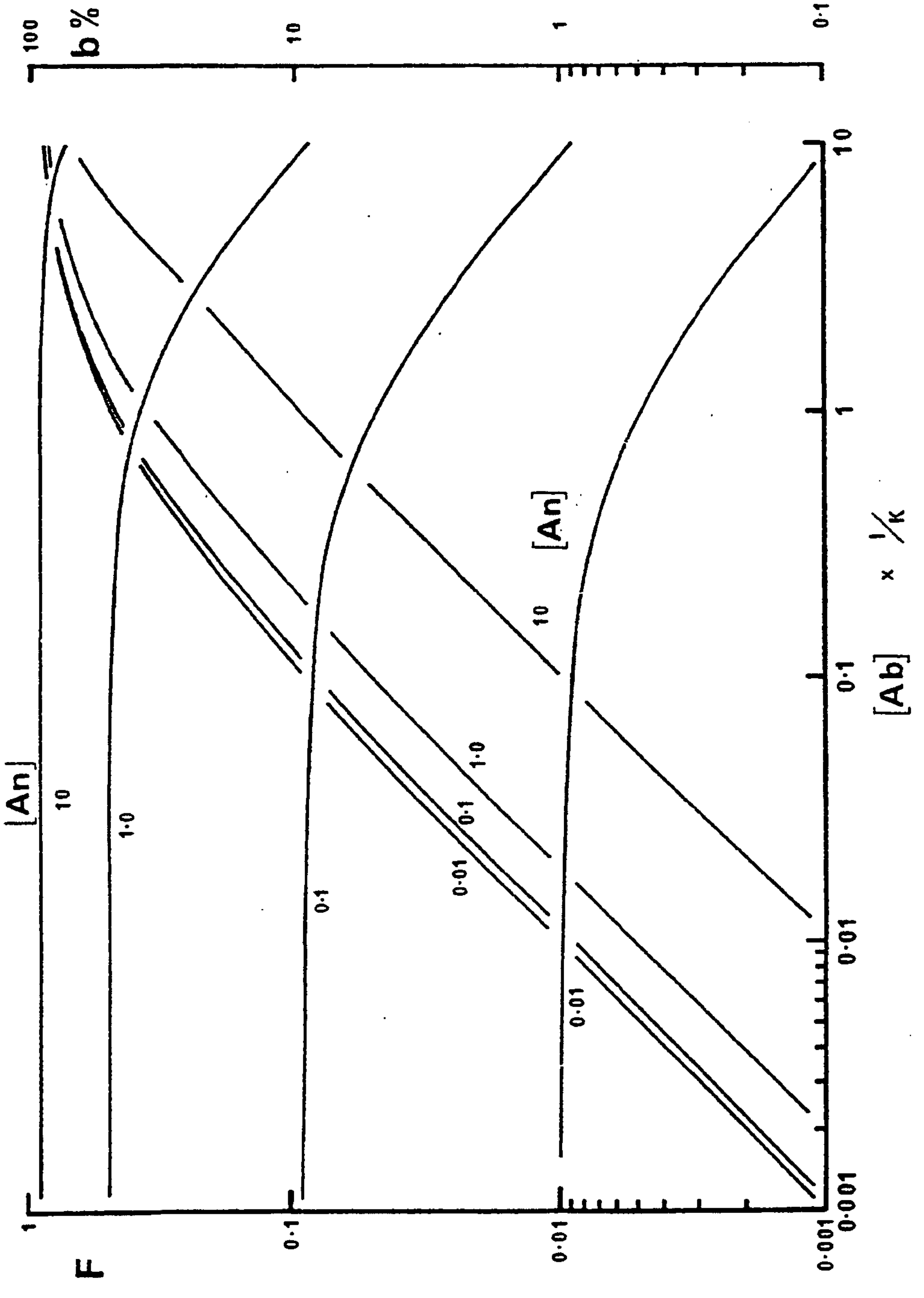
loading a plurality of different binding agents, each being capable of binding specifically and reversibly an analyte of interest onto a support means at a plurality of spaced apart locations such that not more than 0.1 V/K moles of each binding agent are present at any location, where k liters/mole is the equilibrium constant of each such binding agent;

contracting the loaded support means with the sample to be analyzed, such that each of the spaced apart locations is contacted in the same operation with the sample, the amount of sample liquid being such that only an insignificant proportion of any analyte present in the sample becomes bound to the binding agent specific for it, and

measuring a parameter representative of the fractional occupancy by the analytes of the binding agents at the spaced apart locations by a competitive or non-competitive assay technique, using a labelled site-recognition reagent for each binding agent capable of recognizing either the unfilled binding sites or the filled binding sites on the binding agent, which enables the amount of said reagent in the particular location to be measured. A device and kit for use in the method are also provided.

17 Claims, 1 Drawing Sheet




b %
100
10
1
0·1
F
1
0·1
0·01
0·001
[An]
10
1·0
0·1
0·01
[An]
10
1·0
0·1
0·01
[Ab] × 1/K
0·001
0·01
0·1
1
10

DETERMINATION OF AMBIENT CONCENTATION OF SEVERAL ANALYTES

This is a continuation of co-pending application Ser. No. 07/460,878, filed as PCT/GB88/00649, Aug. 5, 1988.

FIELD OF THE INVENTION

The present invention relates to the determination of ambient analyte concentrations in liquids, for example the determination of analytes such as hormones, proteins and other naturally occurring or artificially present substances in biological liquids such as body fluids.

BACKGROUND OF THE INVENTION

I have proposed in International Patent Application WO84/01031 to measure the concentration of an analyte in a fluid by contacting the fluid with a trace amount of a binding agent such as an antibody specific for the analyte in the sense that it reversibly binds the analyte but not other components of the fluid, determining a quantity representative of the proportional occupancy of binding sites on the binding agent and estimating from that quantity the analyte concentration. In that application I point out that, provided that the amount of binding agent is sufficiently low that its introduction into the fluid causes no significant diminution of the concentration of ambient (unbound) analyte, the fractional occupancy of the binding sites on the binding agent by the analyte is effectively independent of the absolute volume of the fluid and of the absolute amount of binding agent, i.e. independent within the limits of error usually associated with the measurement of fractional occupancy. In such circumstances, and in these circumstances only, the initial concentration [H] of analyte in the fluid is related to the fraction (Ab/Ab_o) of binding sites on the binding agent occupied by the analyte by the equation:

$$Ab/Ab_o=K_{ab}[H]/1+K_{ab}[H]$$

where K_{ab} (hereinafter referred to as K) is the equilibrium constant for the binding of the analyte to the binding sites and is a constant for a given analyte and binding agent at any one temperature. This constant is generally known as the affinity constant, especially when the binding agent is an antibody, for example a monoclonal antibody.

The concept of using only a trace amount of binding agent is contrary to generally recommended practice in the field of immunoassay and immunometric techniques. For example, in such a well-known work as "Methods in Investigative and Diagnostic Endocrinology", ed. S. A. Berson and R. S. Yalow, 1973 at pages 111–116, it is proposed that in the performance of a competitive immunoassay maximum sensitivity of the assay is achieved if the proportion of the "tracer" analyte that is bound approximates to 50%. In order to achieve such a high degree of binding of the analyte the theory of Berson and Yalow, to this day generally accepted by other workers in the field, requires that the concentration of binding agent (or, strictly speaking, of binding sites, each molecule of binding agent conventionally having one or at most two binding sites) must be greater than or equal to the reciprocal of the equilibrium constant (K) of the binding agent for the analyte, i.e. [ab]>1/K. For a sample of volume V the total amount of binding agent (or binding sites) must therefore be greater than or equal to V/K. A binding agent which is a monoclonal antibody may, for example, have an equilibrium constant (K) which is of the order of 10^{11} liters/mole for the specific antigen to which it binds. Thus, under the above generally accepted practice, a binding agent (or site) concentration of the order of 10^{-11} mole/liter or more is required for binding agents of such an equilibrium constant and, with fluid sample volumes of the order of 1 milliliter, the use of 10^{-14} or more mole of binding agent (or site) is conventionally deemed necessary. Avogadro's number is about 6×10^{23} so that 10^{-14} mole of binding site is equivalent to more than 10^9 molecules of binding agent even assuming that the binding agent possesses two binding sites per molecule. For specific binding agents of the very highest affinity K is less than 10^{13} liters/mole so that conventional practice requires more than 10^7 molecules of binding agent, whereas binding agents with lower affinity of the order of 10^8 liters/mole necessitate the use of more than 10^{12} molecules under conventional practice. In fact all immunoassay kits marketed commercially at the present time conform to these concepts and use an amount of binding site approximating to or, more frequently, considerably in excess of V/K; indeed in certain types of kit relying on the use of labelled antibodies it is conventional to use as much binding agent as possible, binding proportions of analyte greatly exceeding 50%.

Because of the binding of substantial proportions, for example 50%, of the analyte in the liquid samples under test in such systems, the fractional occupancy of the binding sites of the binding agent is not independent of the volume of the fluid sample so that for accurate quantitative assays it is necessary to control accurately the volume of the sample, keeping it constant in all tests, whether of the sample of unknown concentration or of the standard samples of known concentration used to generate the dose response curve. Furthermore, such systems also require careful control of the amount of binding agent present in the standard and control incubation tubes. These limitations of present techniques are universally recognised and accepted.

UK Patent Application 2,099,578A discloses a device for immunoassays comprising a porous solid support to which antigens, or less frequently immunoglobulins, are bound at a plurality of spaced apart locations, said device permitting a large number of qualitative or quantitative immunoassays to be performed on the same support, for example to establish an antibody profile of a sample of human blood serum. However, although the individual locations may be in the form of so-called microdots produced by supplying droplets of antigen-containing solutions or suspensions, the number of moles of antigen present at each location is apparently still envisaged as being enough to bind essentially all of the analyte (e.g. antibody) whose concentration is to be measured that is present in the liquid sample under test. This is apparent from the fact that the quantitative method used in that application (page 3, lines 21-28) involves calibration with known amounts of immunoglobulin being applied to the support; but this means that, in the samples being tested, essentially every molecule must be extracted from the sample in order for a true comparison to be made and hence that large amounts of antigen (i.e. the binding agent in this situation) are required in each microdot, greatly in excess of the total amount of analyte (i.e. antibody in this situation) present in the sample.

SUMMARY OF THE INVENTION

The present invention involves the realisation that the use of high quantities of binding agent is neither necessary for good sensitivity in immunoassays nor is it generally desirable. If, instead of being kept as large as possible, the amount of binding agent is reduced so that only an insignificant proportion of the analyte is reversibly bound to it, generally less than 10%, usually less than 5% and for optimum results only 1 or 2% or less, not only is it no longer necessary to use an accurately controlled, constant volume for all the liquid samples (standard solutions and unknown samples) in a given assay, but it is also possible to obtain reliable and sometimes even improved estimates of analyte concentration using much less than V/K moles of binding agent binding sites, say not more than 0.1 V/K and preferably less than 0.01 V/K. For a binding agent having an equilibrium constant (K) for the analyte of the order of 10^{11} liters/mole and samples of approximately 1 ml size this is approximately equivalent to not more than 10^8, preferably less than 10^7, molecules of binding agent at each location in an individual array. If the value of K is 10^{13} liters/mole the figures are 10^6 and 10^5 molecules respectively, and if K is of the order of 10^8 liters/mole they are 10^{11} and 10^{10} molecules respectively. Below 10^2 molecules of binding agent at a single location the accuracy of the measurement would become progressively less as the fractional occupancy of the binding agent sites by the analyte would be able to change only in discrete steps as individual sites become occupied or unoccupied, but in principle at least the use of as low as 10 molecules would be permissible if an estimate with an accuracy of 10% is acceptable. Practical considerations may give rise to a preference for more than 10^4 molecules.

It will be appreciated that the abovementioned GB patent application 2,099,578A, which for quantitative estimation relies on large amounts of binding agent and essentially total sequestration of all analyte, fails to recognise the advance achieved by the present invention, which instead relies on a different analytical principle requiring measurement of the fractional occupancy of the binding agent and which thus requires only a very low proportion of the total analyte molecules present to be sequestered from the sample.

Following the recognition that the use of such small amounts of binding agent is permissible, it becomes feasible to place the binding agent required for a single concentration measurement on a very small area of a solid support and hence to place in juxtaposition to one another but at spatially separate points on a single solid support a wide variety of different binding agents specific for different analytes which are or may be present simultaneously in a liquid to be analysed. Simultaneous exposure of each of the separate points to the liquid to be analysed will cause each binding agent spot to take up the analyte for which it is specific to an extent (i.e. fractional binding site occupancy) representative of the analyte concentration in the liquid, provided only that the volume of solution and the analyte concentration therein are large enough that only an insignificant fraction (generally less than 10%, usually less than 5%) of the analyte is bound to the point. The fractional binding site occupancy for each binding agent can then be determined using separate site-recognition reagents which recognise either the unfilled binding sites or filled binding sites of the different binding agents and which are labelled with markers enabling the concentration levels of the separate reagents bound to the different binding agents to be measured, for example fluorescent markers. Such measurements may be performed consecutively, for example using a laser which scans across the support, or simultaneously, for example using a photographic plate, depending on the nature of the labels. Other imaging devices such as a television camera can also be used where appropriate. Because the binding agents are spatially separate from one another it is possible to use only a small number of different marker labels or even the same marker label throughout and to scan each binding agent location separately to determine the presence and concentration of the label. By use of the invention considerably more than 3 analyses can be performed with a single exposure of the solid support with liquid to be analysed, for example 10, 20, 30, 50 or even up to 100 or several hundreds of analyses.

Overall, therefore, the present invention provides a method for determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising:

loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid and is specific for that analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart locations such that each location has not more than 0.1 V/K moles of a single binding agent, where K liters/mole is the equilibrium constant of the binding agent for the analyte,

contacting the loaded support means with the liquid sample to be analysed such that each of the spaced apart locations is contacted in the same operation with the liquid sample, the amount of liquid used in the sample being such that only an insignificant proportion of any analyte present in the liquid sample becomes bound to the binding agent specific for it, and

measuring a parameter representative of the fractional occupancy by the analytes of the binding agents at the spaced apart locations by a competitive or non-competitive assay technique using a site-recognition reagent for each binding agent capable of recognising either the unfilled binding sites or the filled binding sites on the binding agent, said site-recognition reagent being labelled with a marker enabling the amount of said reagent in the particular location to be measured.

The invention also provides a device for use in determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising a solid support means having located thereon at a plurality of spaced apart locations a plurality of different binding agents, each binding agent being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for that analyte as compared to the other components of the liquid sample, each location having not more than 0.1 V/K, preferably less than 0.01 V/K, moles of a single binding agent, where K liters/mole is the equilibrium constant of that binding agent for reaction with the analyte to which it is specific.

A kit for use in the method according to the invention comprises a device according to the invention, a plurality of standard samples containing known concentrations of the analytes whose concentrations in the liquid sample are to be measured and a set of labelled site-recognition reagents for reaction with filled or unfilled binding sites on the binding agents.

In arriving at the method of the invention, I have found that, generally speaking, for antibodies having an affinity constant K liters/mole for an antigen, the relationship between the antibody concentration and the fractional occupancy of the binding sites at any particular antigen concentration and the relationship between the antibody concentration and the percentage of antigen bound to the binding sites at any particular antigen concentration follow the same curves provided that the antibody concentrations and the antigen concentrations are each expressed in terms of fractions or multiples of 1/K.

BRIEF DESCRIPTION OF THE DRAWINGS

The principle underlying the method of the invention may be better understood by reference to the accompanying drawing which is a graph representing two sets of curves plotting the relationship between antibody concentration and the fractional occupancy of the binding sites at certain prescribed antigen concentrations and the relationship between antibody concentration and the percentage of antigen bound to the binding sites at the same prescribed antigen concentrations. Each curve relates to the antibody concentration [Ab], expressed in terms of 1/K, plotted along the x-axis. For the set of curves which remain constant or decline with increasing [Ab], the y-axis represents the fractional occupancy (F) of binding sites on the antibody by the antigen; for the second set, the y-axis represents the percentage (b%) of antigen bound to those Binding sites. The individual curves in each set represent the relationships corresponding to four different antigen concentrations [Ann] expressed in terms of K, namely 10/K, 1.0/K, O.1/K and 0.01/K. The curve show that as [Ab] falls F reaches an essentially constant level, the value of which is dependent on [An].

DETAILED DESCRIPTION

The choice of a solid support is a matter to be left to the user. Preferably the support is non-porous so that the binding agent is disposed on its surface, for example as a monolayer. Use of a porous support may cause the binding agent, depending on its molecular size, to be carried down into the pores of the support where its exposure to the analyte whose concentration is to be determined may likewise be affected by the geometry of the pores, so that a false reading may be obtained. Porous supports such as nitrocellulose paper dotted with spots of binding agent are therefore less preferred, Unlike the supports used in GB 2,099,578A, which seem to need to be porous because of the large number of molecules to be attached, the supports for use in the present invention use much smaller quantities and therefore need not be porous. The non-porous supports may, for example be of plastics material or glass, and any convenient rigid plastics material may be used, Polystyrene is a preferred plastics material, although other polyolefins or acrylic or vinyl polymers could likewise be used.

The support means may comprise microbeads, e.g. of such a plastics material, which can be coated with uniform layers of binding agent and retained in specified locations, e.g. hollows, on a support plate, Alternatively the material may be in the form of a sheet or plate which is spotted with an array of dots of binding agent, It can be advantageous for the configuration of the support means to be such that liquid samples of approximately the volume V liters are readily retained in contact with the plurality of spaced apart locations marked with the different binding agents, For example, the spaced apart locations may be arranged in a well in the support means, and a plurality of wells, each provided with the same group of different binding agents in spaced apart locations, can be linked together to form a microliter plate for use with a plurality of samples.

When the support means is to be used in conjunction with a measuring system involving light scanning, the material, e.g. plastics, for the support is desirably opaque to light, for example it may be filled with an opacifying material which may inter alia be white or black, such as carbon black, when the signals to be measured from the binding agent or the site-recognition reagent are light signals, as from fluorescent or luminescent markers. In general, reflective materials are preferred in this case to enhance light collection in the detecting instrument or photographic plate. The final choice of optimum material is governed by its ability to attach the binding agent to its surface, its absence of background signal emission and its possession of other properties tending to maximise the signal/noise ratio for the particular marker or markers attached to the binding agent situated on its surface. Very satisfactory results have been obtained in the Examples described below by the use of a white opaque polystyrene microliter plate commercially available from Dynatech under the trade name White Microfluor microliter wells.

The binding agents used may be binding agents of different specificity, that is to say agents which are specific to different analytes, or two or more of them may be binding agents of the same specificity but of different affinity, that is to say agents which are specific to the same analyte but have different equilibrium constants K for reaction with it. The latter alternative is particularly useful where the concentration of analyte to be assayed in the unknown sample can vary over considerable ranges, for example 2 or 3 orders of magnitude, as in the case of HCG measurement in urine of pregnant women, where it can vary from 0.1 to 100 or more IU/ml.

The binding agents used will preferably be antibodies, more preferably monoclonal antibodies. Monoclonal antibodies to a wide variety of ingredients of biological fluids are commercially available or may be made by known techniques. The antibodies used may display conventional affinity constants, for example from 10^8 or 10^9 liters/mole upwards, e.g. of the order of 10^{10} or 10^{11} liters/mole, but high affinity antibodies with affinity constants of 10^{12}–10^{13} liters/mole can also be used. The invention can be used with such binding agents which are not themselves labelled. However, it is also possible and frequently desirable to use labelled binding agents so that the system binding agent-/analyte/site-recognition reagent includes two different labels of the same type, e.g. fluorescent, chemiluminescent, enzyme or radioisotopic, one on the binding agent and one on the site-recognition reagent. The measuring operation then measures the ratio of the intensity of the two signals and thus eliminates the need to place the same amount of labelled binding agent on the support when measuring signals from standard samples for calibration purposes as when measuring signals from the unknown samples. Because the system depends solely on measurement of a ratio representative of binding site occupancy, there is also no need to measure the signal from the entire spot but scanning only a portion is sufficient. Each binding agent is preferably labelled with the same label but different labels can be used.

The binding agents may be applied to the support in any of the ways known or conventionally used for coating binding agents onto supports such as tubes, for example by contacting each spaced apart location on the support with a solution of the binding agent in the form of a small drop, e.g. 0.5 microliter, on a 1 mm^2 spot, and allowing them to remain in contact for a period of time before washing the drops away. A roughly constant small fraction of the binding agent present in the drop becomes adsorbed onto the support as a result of this procedure. It is to be noted that the coating density of binding agent on the microspot does not need to be less than the coating density in conventional antibody-coated tubes; the reduction in the number of molecules on each spot may be achieved solely by reduction of the size of the spot rather than the coating density. A high coating density is generally desirable to maximise signal/noise ratios. The sizes of the spots are advantageously less than 10 mm^2 preferably less than 1 mm^2. The separation is desirably, but not necessarily, 2 or 3 times the radius of the spot, or more. These suggested geometries can nevertheless be changed as required, being subject solely to the limitations on the number of binding agent molecules in each spot, the minimum volume of the sample to which the array of spots will be exposed and the means locally available for conveniently preparing an array of spots in the manner described.

Once the binding agents have been coated onto the support it is conventional practice to wash the support, in the case of antibodies as binding agents, with a solution containing albumen or other protein to saturate all remaining non-specific adsorption sites on the support and elsewhere. To confirm that the amount of binding agent in an individual spot will be less than the maximum amount (0.1 V/K) required to conform to the principle of the present invention, the amount of binding agent present on any individual site can be checked by labelling the binding agent with a detectable marker of known specific activity (i.e. known amount of marker per unit weight of binding agent) and measuring the amount of marker present. Thus, if the use of labelled binder is not desired on the solid support used in the method of the invention the binding agent can nevertheless be labelled in a trial experiment and identical conditions to those found in that trial to give rise to correct loadings of binding agent can be used to apply unlabelled binding agent to the supports to be actually used.

The minimum size of the liquid sample (V liters) is correlated with the number of mole of binding agent (less than 0.1 V/K) so that only an insignificant proportion of the analyte present in the liquid sample becomes bound to the binding agent. This proportion is as a general rule less than 10%, usually less than 5% and desirably 1 or 2% or less, depending on the accuracy desired for the assay (greater accuracy being obtained, other things being equal, when smaller pro portions of analyte are bound) and the magnitude of other error-introducing factors present. Sample sizes of the order of one or a few ml or less, e.g. down to 100 microliters or less, are often preferred, but circumstances may arise when larger volumes are more conveniently assayed, and the geometry may be adjusted accordingly. The sample may be used at its natural concentration level or if desired it may be diluted to a known extent.

The site-recognition reagents used in the method according to the invention may themselves be antibodies, e.g. monoclonal antibodies, and may be anti-idiotypic or anti-analyte antibodies, the latter recognising occupied sites. Alternatively, for example for analytes of small molecular size such as thyroxine (T4), unoccupied sites may be recognised using either the analyte itself, appropriately labelled, or the analyte covalently coupled to another molecule—e.g. a protein molecule—which is directly or indirectly labelled. The site-recognition reagents may be labelled directly or indirectly with conventional fluorescent labels such as fluorescein, rhodamine or Texas Red or materials usable in time-resolved pulsed fluorescence such as europium and other lanthanide chelates, in a conventional manner. Other labels such as chemiluminescent, enzyme or radioisotopic labels may be used if appropriate. Each site-recognition reagent is preferably labelled with the same label but different labels can be used in different reagents. The site-recognition reagents may be specific for a single one of the binding agent/analyte spots in each group of spots or in certain circumstances, as with glycoprotein hormones such as HCG and FSH which have a common binding site, they may be cross-reacting reagents able to react with occupied binding sites in more than one of the spots.

In the assay technique the signals representative of the fractional occupancy of the binding agent in the test samples of unknown concentrations of the analytes can be calibrated by reference to dose response curves obtained from standard samples containing known concentrations of the same analytes. Such standard samples need not contain all the analytes together, provided that each of the analytes is present in some of the standard samples. Fractional occupancy may be measured by estimating occupied binding sites (as with an anti-analyte antibody) or unoccupied binding sites (as with an anti-idiotypic antibody), as one is the converse of the other. For greater accuracy it is desirable to measure the fraction which is closer to zero because a change in fractional occupancy of 0.01 is proportionately greater in this case, although for fractional occupancies in the range 25–75% either alternative is generally satisfactory.

In that embodiment of the present invention which relies on two fluorescent markers, the measurement of relative intensity of the signals from the two markers, one on the binding agent and the other on the site recognition reagent, may be carried out by a laser scanning confocal microscope such as a Bio-Rad Lasersharp MRC 500, available from Bio-Rad Laboratories Ltd., and having a dual channel detection system. This instrument relies on a laser beam to scan the dots or the like on the support to cause fluorescence of the markers and wavelength filters to distinguish and measure the amounts of fluorescence emitted. Time-resolved fluorescence methods may also be used. Interference (so-called crosstalk) between the two channels can be compensated for by standard corrections if it occurs or conventional efforts can be made to reduce it. Discrimination of the two fluorescent signals emitted by the dual-labelled spots is accomplished in the present form of this instrument, by filters capable of distinguishing the characteristic wavelength of the two fluorescent emissions; however, fluorescent substances may be distinguished by other physical characteristics such as differing fluorescence decay times, bleaching times, etc., and any of these means may be used, either alone or in combination, to differentiate between two fluorophores and hence permit measurement of the ratio of two fluorescent labelled entities (binding agent and site-recognition reagent) present on an individual spot, using techniques well known in the fluorescence measurement field. When only one fluorescent label is present the same techniques may be used, provided that care is taken to scan the entire spot in each case and the spots contain essentially the same amount of binding agent from one assay to the next when the unknown and standard samples are used.

In the case of other labels, such as radioisotopic labels, chemiluminescent labels or enzyme labels, analogous means of distinguishing the individual signals from one or from each of a pair of such labels are also well known, For example two radioisotopes such as ^{125}I and ^{131}I may be readily distinguished on the basis of the differing energies of their respective radioactive emissions. Likewise it is possible to identify the products of two enzyme reactions, deriving from dual enzyme-labelled antibody couplets, these being e.g. of different colours, or two chemiluminescent reactions, e.g. of different chemiluminescent lifetime or wavelength of light emission, by techniques well known in the respective fields.

The invention may be used for the assaying of analytes present in biological fluids, for example human body fluids such as blood, serum, saliva or Urine. They may be used for the assaying of a wide variety of hormones, proteins, enzymes or other analytes which are either present naturally in the liquid sample or may be present artificially such as drugs, poisons or the like.

For example, the invention may be used to provide a device for quantitatively assaying a variety of hormones relating to pregnancy and reproduction, such as FSH, LH, HCG, prolactin and steroid hormones (e.g. progesterone, estradiol, testosterone and androstene-dione), or hormones of the adrenal pituitary axis, such as cortisol, ACTH and aldosterone, or thyroid-related hormones, such as T4, T3, and TSH and their binding protein TBG, or viruses such as hepatitis, AIDS or herpes virus, or bacteria, such as staphylococci, streptococci, pneumococci, gonococci and enterococci, or tumour-related peptides such as AFP or CEA, or drugs such as those banned as illicit improvers of athletes' performance, or food contaminants. In each case the binding agents used will be specific for the analytes to be assayed (as compared with others in the sample) and may be monoclonal antibodies therefor.

Further details on the methodology are to be found in my International Patent Publication W088/01058, the contents of which are incorporated herein by reference.

The invention is illustrated by the following Examples.

EXAMPLE 1

An anti-TNF (tumour necrosis factor) antibody having an affinity constant for TNF at 25° C. of about 1×10^9 liters/mole is labelled with Texas Red. A solution of the antibody at a concentration of 80 micrograms/ml is formed and 0.5 microliter aliquots of this solution are added in the form of droplets one to each well of a Dynatech Microfluor (opaque white) filled polystyrene microliter plate having 12 wells.

An anti-HCG (human chorionic gonadotropin) antibody having an affinity constant for HCG at 25° C. of about 6×10^8 liters/mole is also labelled with Texas Red. A solution of the antibody at a concentration of 80 micrograms/ml is formed and 0.5 microliter aliquots of this solution are added in the form of droplets one to each well of the same Dynatech Microfluor microliter plate.

After addition of the droplets the plate is left for a few hours in a humid atmosphere to prevent evaporation of the droplets. During this time some of the antibody molecules in the droplets become adsorbed onto the plate. Next, the wells are washed several times with a phosphate buffer and then they are filled with about 400 microliters of a 1% albumen solution and left for several hours to saturate the residual binding sites in the wells. Thereafter they are washed again with phosphate buffer.

The resulting plate has in each of its wells two spots each of area approximately 1 mm^2. Measurement of the amount of fluorescence shows that in each well one spot contains about 5×10^9 molecules of anti-TNF antibody and the other contains about 5×10^9 molecules of anti-HCG antibody. The wells are designed for use with liquid samples of volume 400 microliters, so that 0.1 V/K is 4×10^{-14} moles (equivalent to 2.4×10^{10} molecules) for the anti-TNF antibody and 7×10^{-14} moles (equivalent to 4×10^{10} molecules) for the anti-HCG antibody.

EXAMPLE 2

A microliter plate prepared as described in Example 1 is used in an assay for an artificially produced solution containing TNF and HCG. A test sample of the solution, amounting to about 400 microliters, is added to one of the wells and allowed to incubate for several hours. About 400 microliters of various standard solutions containing known concentrations (0.02, 0.2, 2 and 20 ng/ml) of TNF or HCG are added to other wells of the plate and also allowed to incubate for several hours. The wells are then washed several times with buffer solution.

As site-recognition reagents there are used for the TNF spots an anti-TNF antibody having an affinity constant for TNF at 25° C. of about 1×10^{10} liters/mole and for the HCG spots an anti-HCG antibody having an affinity constant for HCG at 25° C. of about 1×10^{11} liters/mole. Both antibodies are labelled with fluorescein (FITC). 400 microliter aliquots of solutions of these labelled antibodies are added to the wells and allowed to stand for a few hours. The wells are then washed with buffer.

The resulting fluorescence ratio of each spot is quantified with a Bio-Rad Lasersharp MRC 500 confocal microscope. From the standard solutions dose response curves for TNF and HCG are built up, the figures for TNF being as follows:

TNF concentration ng/ml	$\frac{\text{FITC fluorescence}}{\text{Texas Red fluorescence}}$ on TNF spot
0.02	1.1
0.2	4.6
2	7.9
20	42.5

and those for HCG being as follows:

HCG concentration ng/ml	$\frac{\text{FITC fluorescence}}{\text{Texas Red fluorescence}}$ on HCG spot
0.02	1.8
0.2	7.2

-continued

HCG concentration ng/ml	$\frac{\text{FITC fluorescence}}{\text{Texas Red fluorescence}}$ on HCG spot
2	16.0
20	28.2

The artificially produced solution was found to give ratio readings of 5.9 on the TNF spot and 10.5 on the HCG spot, correlating well with the actual concentrations of TNF (0.5 ng/ml) and HCG (0.5 ng/ml) obtained from the dose response curves.

EXAMPLE 3

Using similar procedures to those outlined in Example 1 a microliter plate containing spots of labelled anti-T4 (thyroxine) antibody (affinity constant about 1×10^{11} liters/mole at 25° C), labelled anti-TSH (thyroid stimulating hormone) antibody (affinity constant about 5×10^{9} liters/mole at 25° C.) and labelled anti-T3 (triiodothyronine) antibody (affinity constant about 1×10^{11} liters/mole at 25° C.) in each of the individual wells is produced, the spots containing less than 1×10^{-12} V moles of anti-T4 antibody or less than 2×10^{-11} V moles of anti-TSH antibody or less than 1×10^{-12} V moles of anti-T3 antibody.

The developing antibody (site-recognition reagent) for the TSH assay is an anti-TSH antibody with an affinity constant for TSH of 2×10^{10} liters/mole at 25° C. This antibody is labelled with fluorescein (FITC). The site-recognition reagents for the T4 and T3 assays are T4 and T3 coupled to poly-lysine and labelled with FITC, and they recognise the unfilled sites on their respective first antibodies.

Using 400 microliter aliquots of standard solutions containing various known amounts of T4, T3 and TSH, dose response curves are obtained by methods analogous to those in Example 2, correlating fluorescence ratios with T4, T3 and TSH concentrations. The plate is used to measure T4, T3 and TSH levels in serum from human patients with good correlation with the results obtained by other methods.

EXAMPLE 4

Using similar procedures to those outlined in Example 1 a microliter plate containing spots of first labelled anti-HCG antibody (affinity constant about 6×10^{8} liters/mole at 25° C.), second labelled anti-HCG antibody (affinity constant about 1.3×10^{11} liters/mole at 25° C.) and labelled anti-FSH (follicle stimulating hormone) antibody (affinity constant about 1.3×10^{8} liters/mole at 25° C.) in each of the individual wells is produced, the spots each containing less than 0.1 V/K moles of the respective antibody. A cross-reacting (alpha subunit) monoclonal antibody 8D10 with an affinity constant of 1×10^{11} liters/mole is used as a common developing antibody for both the HCG and the FSH assays.

Using 400 microliter aliquots of standard solutions containing various known concentrations of HCG and FSH, dose response curves are obtained by methods analogous to those in Example 2, correlating fluorescence ratios with HCG and FSH concentrations, the curve obtained with the higher affinity anti-HCG antibody giving more concentration-sensitive results at the lower HCG concentrations whereas the curve from the lower affinity anti-HCG antibody is more concentration-sensitive at the higher HCG concentrations. The plate is used to measure HCG and FSH concentrations in the urine of women in pregnancy testing, giving good correlations with results obtained by other means and achieving effective concentration measurements for HCG over a concentration range of two or three orders of magnitude by correct choice of the best HCG spot and dose response curve.

Production of labelled antibodies

The labelling of the antibodies with fluorescent labels can be carried out by a well known and standard technique, see Leslie Hudson and Frank C. Hay, "Practical Immunology", Blackwell Scientific Publications (1980), pages 11-13, for example as follows:

The monoclonal antibody anti-FSH 3G3, an FSH specific (beta subunit) antibody having an affinity constant (K) of 1.3×10^{8} liters per mole, was produced in the Middlesex Hospital Medical School, and was labelled with TRITC (rhodamine isothiocyanate) or Texas Red, giving a red fluorescence.

The monoclonal antibody anti-FSH 8D10, a cross-reacting (alpha subunit) antibody having an affinity constant (K) of 1×10^{11} liters per mole, was likewise produced in the Middlesex Hospital Medical School and was labelled with FITC (fluorescein isothiocyanate), giving a yellow-green fluorescence.

The general procedure used involved ascites fluid purification (ammonium sulphate precipitation and T-gel chromatography) followed by labelling, according to the following steps:

1.a. Ammonium sulphate purification

1. Add 4.1 ml saturated ammonium sulphate solution to 5 ml anti body preparation (culture supernatant or 1:5 diluted ascites fluid) under constant stirring (45% saturation).

2. Continue stirring for 30-90 min. Centrifuge at 2500 rpm for 30 min.

3. Discard the supernatant and dissolve the precipitate in PBS (final volume 5 ml.). Repeat Steps 1 and 2, OR.

4. Add 3.6 ml saturated ammonium sulphate (40% saturation) under constant stirring. Repeat Step 2.

5. Discard the supernatant and dissolve the pellet in the desired buffer.

6. Dialyse overnight in cold against the same buffer (using fresh, boiled-in-d/w dialysis bag).

7. Determine the protein concentration either at A_{280} or by Lowry estimation.

1.b. T-gel Chromatography: (Buffer: 1M Tris-Cl, pH 7.6. Solid potassium sulphate)

1. Clear 2 ml of ascites fluid by centrifugation at 4000 rpm.

2. Add 1M Tris-Cl solution to achieve final concentration of 0.1M.

3. Add sufficient amount of solid potassium sulphate. Final concentration:=0.5M.

4. Apply the ascite fluid to the T-gel column.

5. Wash the column with 0.1M Tris-Cl buffer containing 0.5M potassium sulphate, until protein profile (at A_{280}) returns to zero.

6. Elute the absorbed protein using 0.1M Tris-Cl buffer as the eluant.

7. Pool the fractions containing antibody activity and concentrate using Amicon 30 concentrater.

8. If HPHT purification is to be carried out, use HPHT chromatography Starting buffer during Step 7.

2. Labelling of Antibodies FITC/TRITC conjugation:

1. Dialyse the purified 1 g protein into 0.25M Carbonate-bicarbonate buffer, pH 9.0 to a concentration of 20 mg/ml.

2. Add FITC/TRITC to achieve a 1:20 ratio with protein (i.e. 0.05 mg for every 1 mg of protein).

3. Mix and incubate at 4° C. for 16–18 hrs.

4. Separate the conjugated protein from unconjugated by:

a. Sephadex G-25 chromatography for FITC label, or

b. DEAE-Sephacel chromatography for TRITC-/FITC label.

Buffer system:

PBS for (a).

0.005M Phosphate, pH 8.0 and 0.18M Phosphate, pH 8.0 for (b).

$$\frac{2.87 \times \text{O.D.}495\text{ nm}}{\text{O.D.}280\text{ nm} - 0.35 \times \text{O.D.}495\text{ nm}}$$

I claim:

1. A method for determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising:

loading a plurality of different binding agents, each being capable of reversibly binding an analyte which is or may be present in the liquid sample and is specific for said analyte as compared to the other components of the liquid sample, onto a support means at a plurality of spaced apart small spots such that each spot has a high coating density of one of said binding agents but not more than 0.1 V/K moles of binding agent are present on any spot, where K liters/mole is the affinity constant of said binding agent for said analyte;

contacting the loaded support means with the liquid sample to be analyzed, such that each of the spots is contacted in the same step with said liquid sample, the amount of liquid used in said sample being such that only an insignificant proportion of any analyte present in said liquid sample becomes bound to said binding agent specific for said analyte, and

measuring a parameter representative of the fractional occupancy by said analytes of said binding agents at the spots by a competitive or non-competitive assay technique using a site-recognition reagent for each binding agent capable of recognizing either the unfilled binding sites or the filled binding sites on said binding agent, said site-recognition reagent being labelled with a marker enabling the amount of said reagent in the particular location to be measured.

2. A method as claimed in claim 1, wherein each of said spots has a size of less than 1 mm^2.

3. A method as claimed in claim 2, wherein each of said spots contains more than 10^4 molecules of binding agent.

4. A method as claimed in claim 3, wherein each of said spots has less than 0.01 V/K moles of binding agent.

5. A method as claimed in claim 3, wherein said binding agents used have affinity constants for said analytes of from 10^8 to 10^{13} liters per mole.

6. A method as claimed in claim 3, wherein said binding agents used have affinity constants for said analytes of the order of 10^{10} or 10^{11} liters per mole.

7. A method as claimed in claim 3, wherein the volume of said liquid sample is not more than 0.1 liter.

8. A method as claimed in claim 3, wherein the volume of said liquid sample is 400 to 1000 microliters.

9. A method as claimed in claim 1, wherein said binding agents loaded onto said support means are antibodies for the analytes whose concentrations are to be determined.

10. A method as claimed in claim 1, wherein said binding agents are labelled with markers enabling the concentration levels of said binding agents to be measured.

11. A method as claimed in claim 10, wherein said binding agents and said site-recognition reagents are labelled with fluorescent markers such that at the individual spots the assay technique for measuring fractional occupancy of the binding agents measures the ratios of the signals emitted by the fluorescent markers.

12. A device for use in determining the ambient concentrations of a plurality of analytes in a liquid sample of volume V liters, comprising a solid support means having located thereon at high coating density at a plurality of spaced apart small spots a plurality of different binding agents, each binding agent being capable of reversibly binding an analyte which is or may be present in said liquid sample and is specific for said analyte as compared to the other components of said liquid sample, each spot having not more than 0.1 V/K moles of a single binding agent, where K liters/mole is the affinity constant of said single binding agent for reaction with the analyte to which it is specific.

13. A device as claimed in claim 12, wherein each of said spots has a size of less than 1 m^2.

14. A device as claimed in claim 13, wherein each of said spots contains more than 10^4 molecules of binding agent.

15. A kit for use in determining the ambient concentration of a plurality of analytes in a liquid sample of volume V liters, comprising:

a solid support means having located thereon at high coating density at a plurality of spaced apart small spots a plurality of different binding agents, each binding agent being capable of reversibly binding an analyte which is or may be present in said liquid sample and is specific for said analyte as compared to the other components of the liquid sample, each spot having not more than 0.1 V/K moles of a single binding agent, where K liters/mole is the affinity constant of said single binding agent for reaction with the analyte to which it is specific;

a plurality of standard samples containing known concentrations of the analytes whose concentrations in the liquid sample are to be measured; and

a set of labelled site-recognition reagents for reaction with filled or unfilled binding sites on said binding agents.

16. A kit as claimed in claim 15, wherein each of said spots has a size of less than 1 mm^2.

17. A kit as claimed in claim 16, wherein each of said spots contains more than 10^4 molecules of binding agent.

* * * * *

Exhibit Q of J. C. Rockett Declaration
dated 12/04/03
In USSN 09/937,060

US005837551A

United States Patent [19]

Ekins

[11] **Patent Number: 5,837,551**

[45] **Date of Patent: Nov. 17, 1998**

[54] **BINDING ASSAY**

[76] Inventor: **Roger P. Ekins**, Pondweed Place, Friday Street, Abinger, Common Dorking Surrey, Great Britain, RH5 GJR

[21] Appl. No.: **663,176**

[22] PCT Filed: **Dec. 23, 1994**

[86] PCT No.: **PCT/GB94/02814**

§ 371 Date: **Jun. 14, 1996**

§ 102(c) Date: **Jun. 14, 1996**

[87] PCT Pub. No.: **WO95/18377**

PCT Pub. Date: **Jul. 6, 1995**

[30] **Foreign Application Priority Data**

Dec. 24, 1993 [GB] United Kingdom 9326450

[51] **Int. Cl.**^{6} **G01N 33/53**

[52] **U.S. Cl.** **436/518**; 435/6; 435/7.1; 435/7.92; 435/962; 435/970; 435/973; 435/975; 436/518; 436/809

[58] **Field of Search** 435/6, 7.1, 7.92, 435/962, 970, 973, 975; 436/518, 809

[56] **References Cited**

U.S. PATENT DOCUMENTS

5,096,807	3/1992	Leaback	435/6
5,324,633	6/1994	Fodor et al.	435/6
5,432,099	7/1995	Ekins	436/518
5,508,200	4/1996	Tiffany et al.	436/44
5,552,272	9/1996	Bogart	435/6
5,599,668	2/1997	Stimpson et al.	435/6

FOREIGN PATENT DOCUMENTS

0304202	2/1989	European Pat. Off. .
WO8401031	3/1984	WIPO .
WO8801058	2/1988	WIPO .
WO9308472	4/1993	WIPO .

OTHER PUBLICATIONS

Ekins et al., "Multianalyte Microspot Immunoassay—Microanalytical 'Compact Disk' of the Future," Clinical Chemistry, 37 (11):1955–67, 1991.

Berson and Yalow, "Methods in Investigative and Diagnostic Endocrinology", pp. 111–116 (1973).

Primary Examiner—Carol A. Spiegel
Attorney, Agent, or Firm—Dann, Dorfman, Herrell and Skillman

[57] **ABSTRACT**

The present invention provides methods for determining the concentration of analytes in liquid samples in which the amount of binding agent having binding sites specific for a given analyte in the liquid sample is immobilized in a test zone on a solid support, the binding agent being divided into an array of spatially separated locations in the test zone. The concentration of the analyte is obtained by back-titrating the occupied binding agent with a developing agent having a marker and integrating the signal from each location in the array. The present invention also provides a method for determining a value representative of a fraction of binding sites of the binding agent which are occupied by the analyte, comprising immobilizing the specific binding agent on a solid support, wherein the specific binding agent used for the fractional occupancy is present in an amount less than 0.1 V/K moles, where V is the volume of the liquid sample and K is the association constant for the analyte specifically binding to the binding agent, and wherein the specific binding agent is divided into an array of spatially separated locations; contacting the support with the liquid sample; contacting the support with the developing agent; separating non-specifically bound developing agent and measuring the signal at each of the locations to obtain a value which represents the fraction of the binding sites occupied by the analyte at each location; and adding the measured values to provide a total signal which indicates the fraction of the binding sites of the binding agent occupied by the analyte. Test kits and devices used in practicing these methods are also disclosed.

21 Claims, 3 Drawing Sheets

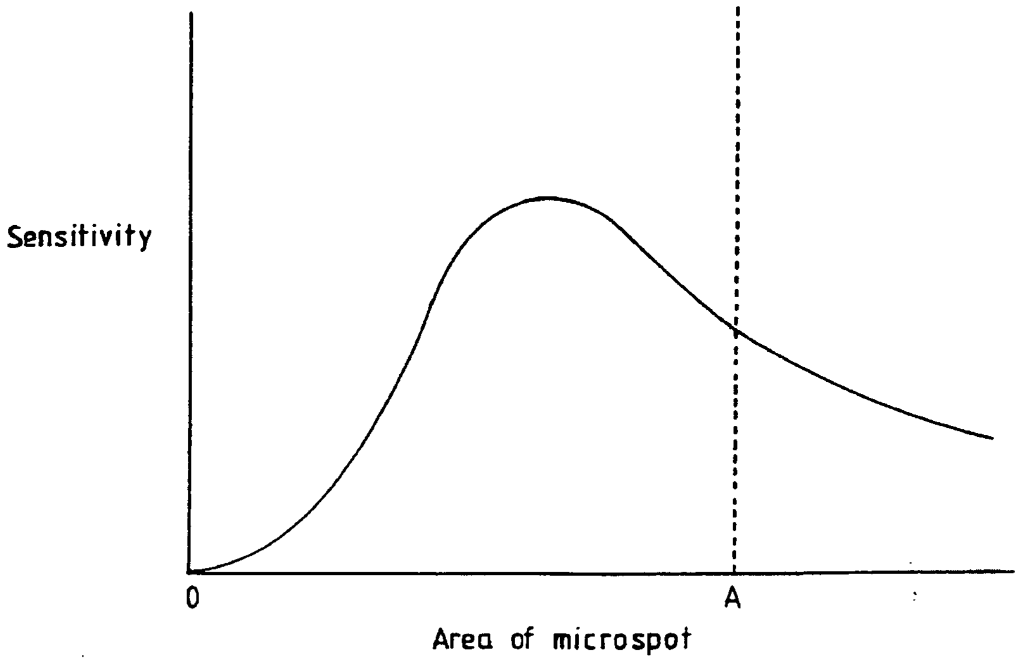

Sensitivity
0
A
Area of microspot

Fig.1.

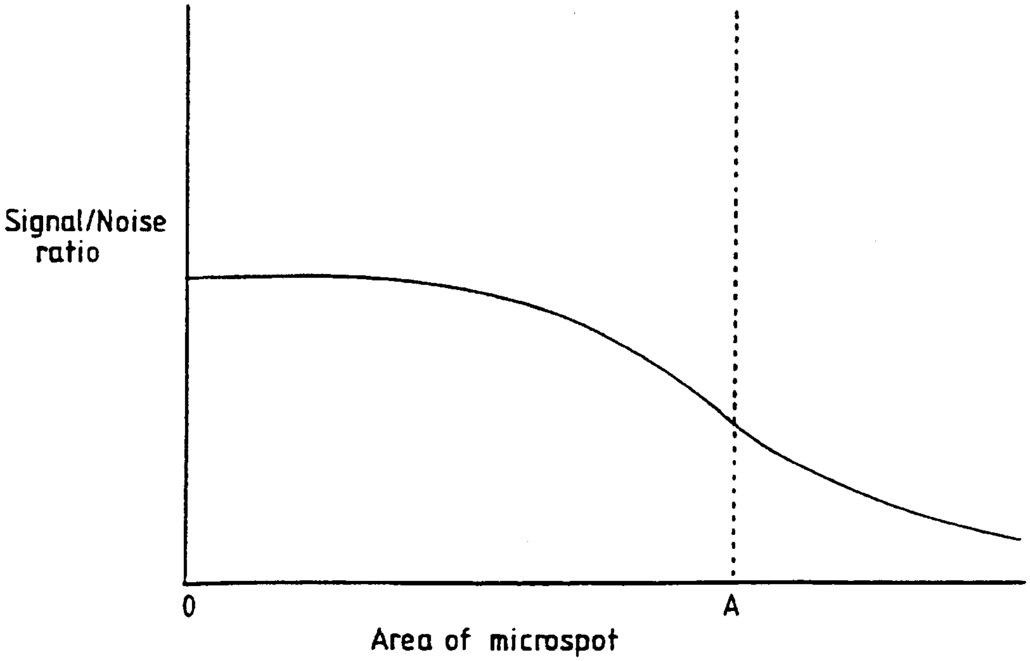

Signal/Noise ratio
0
A
Area of microspot

Fig.2.

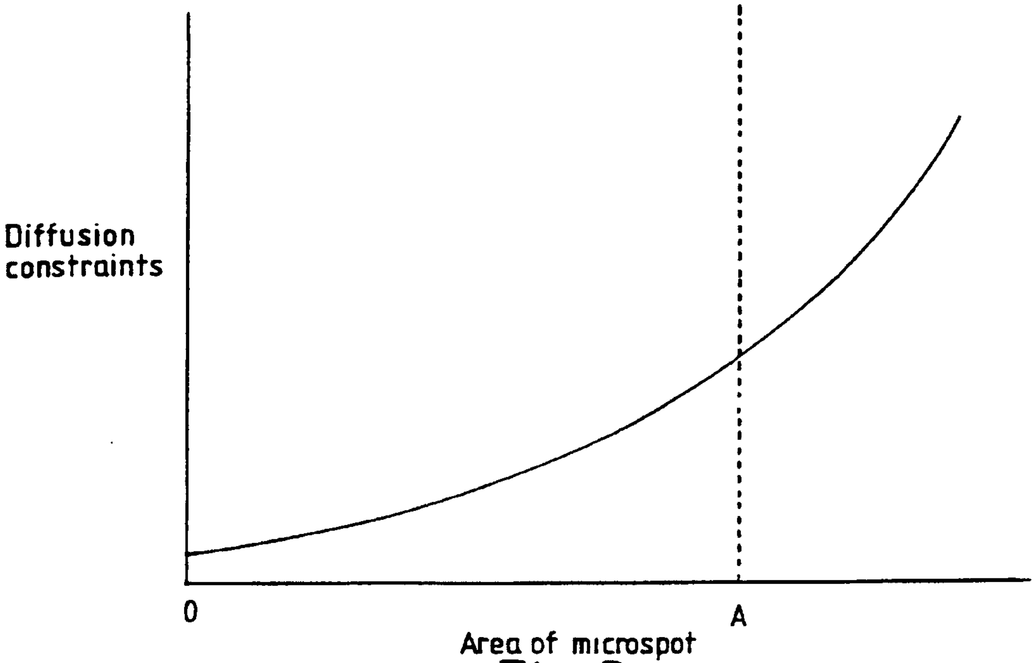

Diffusion constraints
0
A
Area of microspot

Fig.3.

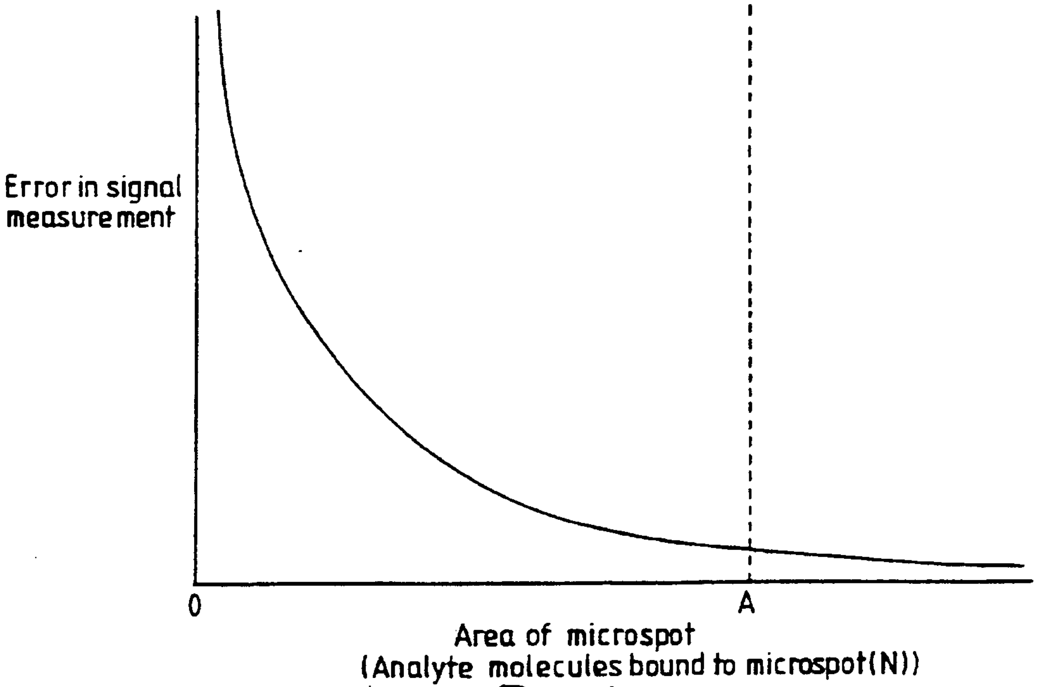

Error in signal measurement
0
A
Area of microspot
(Analyte molecules bound to microspot(N))

Fig.4.

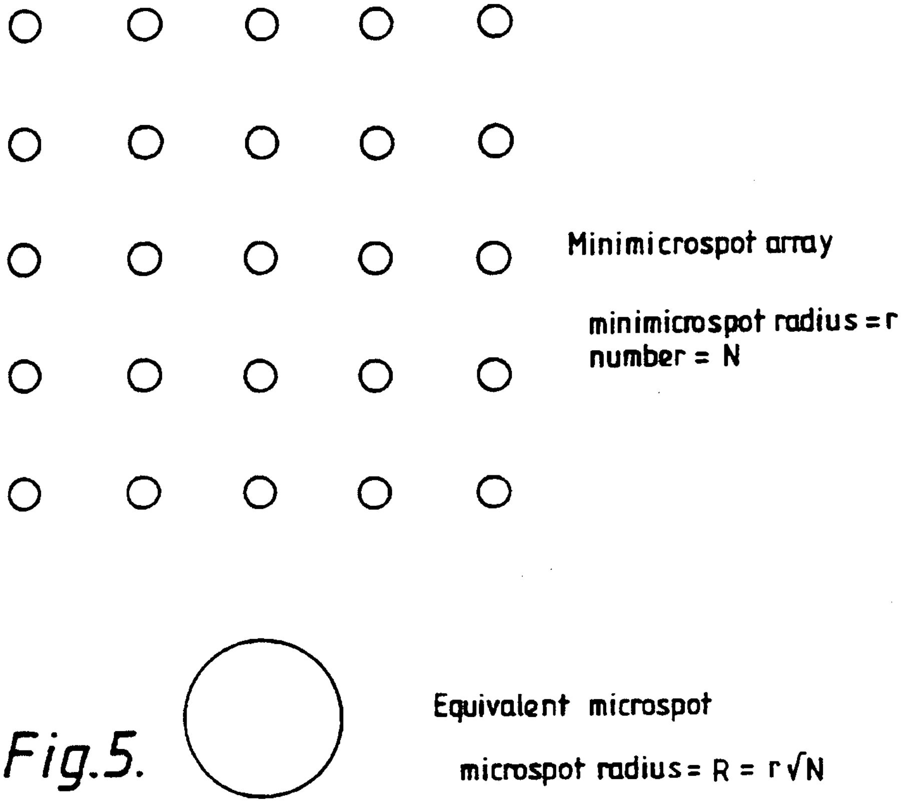

Minimicrospot array
minimicrospot radius = r
number = N
Equivalent microspot
microspot radius = R = r√N

Fig.5.

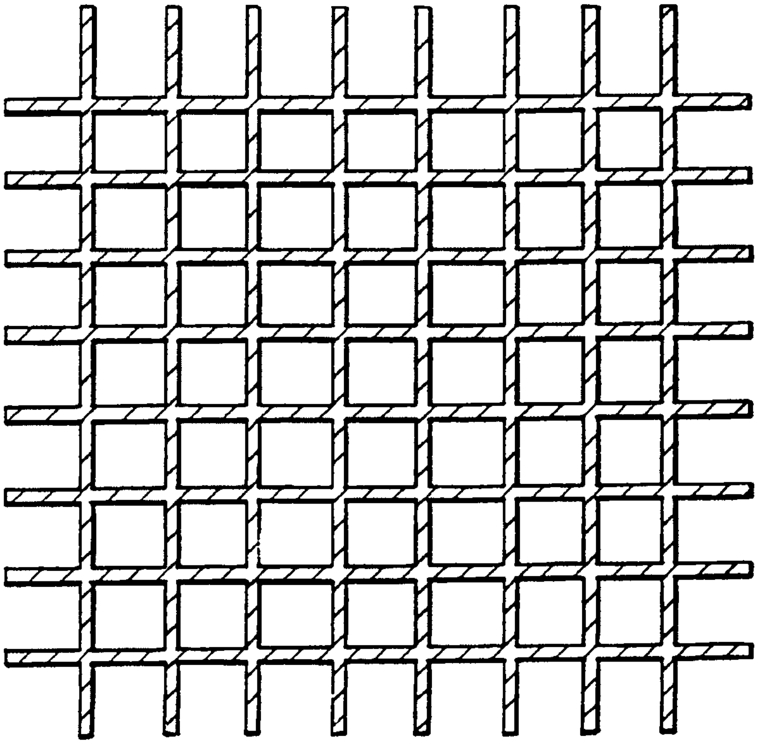


Fig.6.

BINDING ASSAY

This application is the U.S. national stage of PCT/GB94/02814, filed Dec. 23, 1994.

FIELD OF THE INVENTION

The present invention relates to binding assays, e.g. for determining the concentration of analytes in liquid samples.

BACKGROUND TO THE INVENTION

It is known to measure the concentration of an analyte, such as a drug or hormone, in a liquid sample by contacting the liquid with a binding agent having binding sites specific for the analyte, separating the binding agent having analyte bound to it and measuring a value representative of the proportion of the binding sites on the binding agent that are occupied by analyte (referred to as the fractional occupancy). Typically, the concentration of the analyte in the liquid sample can then be determined by comparing the fractional occupancy against values obtained from a series of standard solutions containing known concentrations of analyte.

In the past, the measurement of fractional occupancy has usually been carried out by back-titration with a labelled developing reagent using either so-called competitive or non-competitive methods.

In the competitive method, the binding agent having analyte bound to it is back-titrated, either simultaneously or sequentially, with a labelled developing agent, which is typically a labelled version of the analyte. The developing agent can be said to compete for the binding sites on the binding agent with the analyte whose concentration is being measured. The fraction of the binding sites which become occupied with the labelled analyte can then be related to the concentration of the analyte in the liquid sample as described above.

In the non-competitive method, the binding agent having analyte bound to it is back-titrated with a labelled developing agent capable of binding to either the bound analyte or the occupied binding sites on the binding agent. The fractional occupancy of the binding sites can then be measured by detecting the presence of the labelled developing agent and, just as with competitive assays, related to the concentration of the analyte in the liquid sample as described above.

In both competitive and non-competitive methods, the developing agent is labelled with a marker. A variety of markers have been used in the past, for example radioactive isotopes, enzymes, chemiluminescent markers and fluorescent markers.

In the field of immunoassay, competitive immunoassays have in general been carried out in accordance with design principles enunciated by Berson and Yalow, for instance in "Methods in Investigative and Diagnostic Endocrinology" (1973), pages 111 to 116. Berson and Yalow proposed that in the performance of competitive immunoassays, maximum sensitivity is achieved if an amount of binding agent is used to bind approximately 30 to 50% of a low concentration of the analyte to be detected. In non-competitive immunoassays, maximum sensitivity is generally thought to be achieved by using sufficient binding agent to bind close to 100% of the analyte in the liquid sample. However, in both cases immunoassays designed in accordance with these widely accepted precepts require the volume of the sample to be known and the amount of binding agent used to be accurately known or known to be constant.

In International Patent Application WO84/01031, I disclosed that the concentration of an analyte in a liquid sample can be measured by contacting the liquid sample with a small amount of binding agent having binding sites specific for the analyte. In this method, provided the amount of binding agent is small enough to have only an insignificant effect on the concentration of the analyte in the liquid sample, it is found that the fractional occupancy of the binding sites on the binding agent by the analyte is effectively independent of the volume of the sample.

This approach is further refined in EP304,202 which discloses that the sensitivity and ease of development of the assays in WO84/01031 is improved by using an amount of binding agent less than 0.1V/K moles located on a small area (or "microspot") of a solid support, where V is the volume of the sample and K is the equilibrium constant of the binding agent for the analyte.

In WO93/08472, I disclosed a method of further improving the sensitivity of binding assays by immobilising small amounts of binding agent at high density on a support in the form of a microspot. In this assay, a developing agent comprising a microsphere containing a marker, e.g. a fluorescent dye, is used to back-titrate the binding agent after it has been contacted with the liquid sample containing the analyte. As the microsphere can contain a large number of molecules of fluorescent dye, the sensitivity of the assay is improved as the signal from small amounts of analyte can be amplified. This amplification permits sensitive assays to be carried out even with microspots having an area of 1 mm^2 or less and a surface density of binding agent in the range of 1000 to 100000 molecules/μm^2.

SUMMARY OF THE INVENTION

The present invention provides a method, device and test kit for carrying out a binding assay in which binding agent having binding sites specific for a given analyte in a liquid sample is immobilised in a test zone on a solid support, the binding agent being divided into an array of spatially separated locations in the test zone, wherein the concentration of the analyte is obtained by integrating the signal from the locations in the array.

Accordingly, in one aspect, the present invention provides a method for determining the concentration of an analyte in a liquid sample comprising:

(a) locating binding agent having binding sites specific for the analyte in a test zone on a solid support, the binding agent being divided into an array of spatially separated locations;

(b) contacting the support with the liquid sample so that a fraction of the binding sites at each location become occupied by analyte;

(c) measuring a value of a signal representative of the fraction of the binding sites occupied by the analyte for each individual location in the array;

(d) integrating the signal value obtained for each location in the array to provide an integrated signal; and,

(e) comparing the integrated signal to corresponding values, obtained from a series of standard solutions containing known concentrations of analyte, to determine the concentration of the analyte in the liquid sample.

Thus, in the present invention, the values of the signal from an array of locations in the test zone are used to determine the concentration of a single analyte. This is in contrast to the approach described in EP304,202, in which the signal produced at a single location is used to determine the concentration of an analyte.

The array of locations of binding agent in the test zone can be viewed sequentially, e.g. using a confocal microscope, and the signal value from each location integrated to provide the integrated signal. Alternatively, the array of locations of binding agent in the test zone can be viewed together, e.g. using a charge coupled device (CCD) camera, with the signal values from each location being measured simultaneously.

Preferably, the signals representative of the fraction of the binding sites occupied by binding agent at each location are measured by back-titrating the binding agent with a developing agent having a marker, the developing agent being capable of binding to unoccupied binding sites, bound analyte or to occupied binding sites in a competitive or non-competitive method, as described above.

The marker on the developing agent can be a radioactive isotope, an enzyme, a chemiluminescent marker or a fluorescent marker. The use of fluorescent dye markers is especially preferred as the fluorescent dyes can be selected to provide fluorescence of an appropriate colour range (excitation and emission wavelength) for detection. Fluorescent dyes include coumarin, fluorescein, rhodamine and Texas Red. Fluorescent dye molecules having prolonged fluorescent periods can be used, thereby allowing time-resolved fluorescence to be used to measure the strength of the fluorescent signal after background fluorescence has decayed. Advantageously, marker can be incorporated within or on the surface of latex microspheres attached to the developing agent. This allows a large quantity of marker to be associated with each molecule of developing agent, amplifying the signal produced by the developing agent.

Preferably, the locations are microspots and the assay is carried out using 4–40 (or more) microspots for each individual analyte, each microspot having an area less than 10000 μm^2, the microspots being separated from each other by a distance of 100–1000 μm. The locations within the array are referred to as "mini-microspots" in the relevant parts of the description that follow.

The present invention also allows the concentration of a plurality of analytes to be determined simultaneously by providing a plurality of test zones, each test zone having immobilised in it a total amount of binding agent having binding sites specific for a given analyte in a liquid sample, the binding agent being divided into an array of spatially separated locations in the test zone.

Preferably, in accordance with EP304,202, the total amount of binding agent in each array that is specific for a given analyte is less than 0.1 V/K moles, where V is the volume of the sample applied to the test zone and K is the association constant for analyte binding to the binding agent. This ensures that the "ambient analyte" conditions described in WO83/01031 are fulfilled regardless of the analyte concentration.

One way of immobilising binding agent on a support at a discrete location such as a microspot is to use technology comparable to the techniques used in ink-jet or laser printers, in the case of microspots typically providing spots having diameter of about 80 μm. Alternatively, if larger locations are required, a micropipette can be used to control the amount of binding agent immobilised at a location on a support.

The present invention is based on the observation that as the area of a microspot is reduced from a high value, such as 5 mm^2 towards zero, the sensitivity of the binding assay (represented by the lower limit of detection) reaches a maximum when the microspot reaches a small, but finite area, typically around 0.1 mm^2. Further reducing the area of microspot leads to a reduction in the sensitivity of the binding assay. However, sensitivity is enhanced by subdividing a microspot of any given area into multiple mini-microspots, such that the total coated area occupied by the mini-microspots, and hence the total amount of binding agent remains the same.

In addition, the use of an array of microspots for each individual analyte allows the user to determine whether the value obtained for any given microspot is in error by comparison with other microspots in the array.

In a further aspect, the present invention provides a device for determining the concentration of one or more analytes in a liquid sample, the device comprising a solid support having one or more test zones, each test zone having immobilised in it an amount of binding agent having binding sites specific for a given analyte in a liquid sample, the binding agent being divided into an array of spatially separated locations in the test zone, wherein the concentration of a given analyte is obtained by integrating signal values from each location in the array.

In a further aspect, the present invention provides a kit for determining the concentration of one or more analytes in a liquid sample, the kit comprising:

(a) a device comprising a solid support having one or more test zones, each test zone having immobilised in it an amount of binding agent having binding sites specific for a given analyte in a liquid sample, the binding agent being divided into an array of spatially separated locations in the test zone; and,

(b) one or more developing agents for determining the fraction of the binding sites of the binding agent occupied by a given analyte, the developing agents having markers, the developing agents being capable of binding to bound analyte, or unoccupied or occupied binding sites of the binding agent;

wherein the concentration of a given analyte is obtained by integrating signal values from the markers of the developing agent at each location in the array.

BRIEF DESCRIPTION OF THE DRAWINGS

The unexpected observation of a microspot area yielding a maximum sensitivity is thought to arise because a number of opposing effects combine to produce this outcome. These effects are explained with reference to the accompanying figures in which:

FIG. 1 represents how the sensitivity of a binding assay typically changes with area of microspot at constant binding agent density;

FIG. 2 represents the typical variation in signal-to-noise ratio as the area of a microspot changes;

FIG. 3 represents how diffusion constraints on analyte binding to the binding agent change as the area of the microspot changes;

FIG. 4 shows how the error in signal measurement changes as the area of the microspot changes;

FIG. 5 shows a comparison between the microspot array of the present invention and a single microspot of the prior art; and

FIG. 6 shows binding agent immobilised as an array of lines in an alternative embodiment of the invention.

DETAILED DESCRIPTION

Note that in FIGS. 1 to 4, the exact shape of the curves shown will depend on a number of parameters, including the physico-chemical properties (ie association and dissociation rate constants) of the binding agent, the viscosity of the analyte containing solution to which the microspot is exposed, the specific activity of the label used, etc.

In all the figures, value A denotes the area of a microspot typically used in the prior art (typically 1 mm^2). In all the figures, the density of binding agent is kept constant.

FIG. 1 shows the experimentally observed variation of sensitivity as the area of a microspot is reduced. In the present context, sensitivity can be defined as the lower limit of detection which is given by the error (s.d) with which it is possible to measure zero signal. As FIG. 1 shows, as the area is reduced from value A, the sensitivity of the binding assay reaches a maximum and then declines as the area of the microspot is further reduced towards zero.

Some of the opposing factors leading to this observation are depicted in FIGS. 2 to 4.

FIG. 2 shows how the signal-to-noise ratio associated with the measurement of the occupancy of the binding sites of the binding agent changes as the size of the microspot decreases towards zero, assuming equilibrium has been reached. As microspot area is reduced from value A, the fractional occupancy of the binding sites of the binding agent reaches a plateau value as the concentration of binding agent falls below 0.01/K. Therefore, the signal per unit area from markers on developing agent used to measure the occupancy of the binding sites by analyte will also reach a plateau. As the background noise per unit area remains approximately constant, so the signal-to-noise ratio will likewise increase to a plateau value as the concentration of binding agent falls below 0.01/K.

FIG. 3 shows how diffusion constraints change as the area of a microspot is reduced. "Diffusion constraints" restrict the rate at which analyte migrates towards and binds to the binding agent. As FIG. 3 shows, the diffusion constraints decrease as microspot size decreases, ie the kinetics of the binding process are faster for smaller microspots, implying that thermodynamic equilibrium in the system is reached more rapidly.

On a molecular level, this phenomenon can be pictured as follows. When a microspot containing binding agent is placed in a liquid sample containing analyte, the binding agent binds analyte, depleting the local concentration of the analyte as compared to the liquid sample as a whole. This leads to a concentration gradient being established in the vicinity of the microspot until thermodynamic equilibrium is reached. This process is found to be slower for larger microspots the diffusion constraint being approximately proportional to microspot radius. When the occupancy of the binding sites on the binding agent has reached an equilibrium value, the concentration of analyte in the liquid sample is uniform. However, equilibrium is reached more rapidly in the case of microspots of smaller size, implying that, for any incubation time less than that required to reach equilibrium in the case of the larger spot, the fractional occupancy of the binding sites on the smaller spot is greater.

However, as microspot area decreases, so the amount of binding agent and the level of signal from developing agent will likewise decrease. This leads to an increase in the statistical errors in the measurement of the signal from a marker on a developing agent, which tend to infinity as the microspot area tends to zero (see FIG. 4).

It can be seen that a consideration of the signal-to-noise ratio and diffusion constraints indicate an increase in the sensitivity of a binding assay as the area of a microspot is decreased. However, these factors are opposed by an increase in the statistical error of signal measurement as the microspot area decreases. These factors combine to produce the observed variation of sensitivity with microspot area shown in FIG. 1. Thus, the overall consequence is that, as microspot area falls to zero, the binding assay becomes totally insensitive.

However, it is desirable to develop sensitive miniaturised binding assays using microspots of the smallest possible size containing vanishingly small amounts of binding agent, that have rapid kinetics to minimise the time taken to carry out the assay.

The present invention improves sensitivity and reduces binding assay incubation times by exploiting the contradictory effects discussed above to maximal advantage. This is done by sub-dividing the total amount of binding agent into an array of spatially separated locations such as "minimicrospots", to reduce diffusion constraints, and integrating the signals representative of the fractional occupancy of binding agent at each location to obtain a total signal greater than would have been achieved by using a single microspot equal in area to the total area occupied by the minimicrospots comprising the minimicrospot array.

This implies, inter alia, that the total amount of binding agent used can be made even smaller than in the prior art where a balance between kinetics and signal-to-noise relative to statistical errors had to be made to optimise sensitivity. The present invention therefore can improve the signal-to-noise ratio associated with measuring the analyte bound to binding agent, whilst reducing the diffusion constraints associated with each microspot in the array. Moreover, the increasing statistical errors observed in the prior art as microspot size is reduced are obviated, as the signal generated from the occupied binding sites by analyte in the individual microspots is integrated over the array to provide an integrated signal, thereby retaining the signal measurement advantage observed for larger microspots.

FIG. 5 illustrates how a single microspot of the prior art can be divided into an array of 25 microspots containing an equivalent total amount of binding agent.

Nevertheless, other arrangements or geometries of binding agent providing assays yielding the same benefits can be envisaged, see for instance FIG. 6 which shows binding agent immobilised as lines forming a grid (see the shaded areas). This configuration likewise has the effect of reducing the diffusion constraints whilst maintaining the total area coated with binding agent (e.g. an antibody) to obviate the increasing statistical errors and associated loss of sensitivity observed as the amount of binding agent is reduced.

The amount and distribution of the binding agent in the locations comprising the array depends on a variety of factors including the diffusion characteristics of the analyte, the nature and viscosity of the liquid sample containing the analyte and the protocol used during incubation. However, given the guidance here the skilled person can readily determine, either experimentally or by computer modelling, the optimal arrangement or geometry of array for any given binding assay.

EXAMPLE

Conjugation of Anti-TSH (Anti-Thyroid Stimulating Hormone) Mouse Monoclonal Antibody to Fluorescent Hydrophilic Latex Microspheres

1. 10 mg of fluorescent hydrophilic latex microspheres in 0.5 ml double distilled water were added to 0.5 ml of 1% TWEEN 20, surface-active agent, shaken for 15 min at room temperature and centrifuged at 8° C. for 10 min at 20,000 rpm in a MSE High-Spin 20 ultracentrifuge.

2. The pellet was dispersed in 2 ml of 0.05M MES (2-[N-Morpholino] ethanesulfonic acid) buffer, pH6.1 and centrifuged.

3. Step 2 was repeated.

4. The pellet was dispersed in 0.8 ml MES buffer.

5. 2 mg of anti-TSH monoclonal developing antibody in 100 μl were added to the microspheres and shaken for 15 min at room temperature.

6. 100 μl of 0.25% ethyl-3 (3-dimethyl amino) propyl carbodimide hydrochloride were added to the mixture and shaken for 2 hours at room temperature.

7. 10 mg glycine in 100 μl of MES buffer were added to the mixture, shaken for a further 30 min and centrifuged.

8. The pellet was dispersed in 2 ml of 1% BSA (Bovine Serum Albumin), shaken for 1 hour at room temperature and centrifuged.

9. The pellet was dispersed in 2 ml of 1% BSA, shaken for 1 hour at room temperature and centrifuged.

10. The pellet was dispersed in 2 ml of 0.1M phosphate buffer, pH7.4 and centrifuged.

11. Step 10 was repeated twice.

12. The pellet was dispersed in 2 ml of 1% BSA containing 0.1 sodium azide and stored at 4° C.

Comparison of Kinetics of Micro Versus Mini-micro Capture Antibody Microspots in a Sandwich TSH (Thyroid Stimulating Hormone) Assay

1. Anti-TSH capture antibody microspots (diameter 1.1 mm, area=10^6 μm^2) were made by depositing 0.5 μl of 200 μg/ml antibody solution on each of 16 Dynatech black MicroFluor Microtitre wells, the droplets were aspirated immediately, the wells blocked with SuperBlock from Pierce for 30 min at room temperature and washed with 0.1M phosphate buffer, pH7.4.

2. The mini-microspots (diameter 0.16 mm) were made using an piezoelectric ink-jet print-head with an anti-body solution concentration of 1 mg/ml and droplets of approximately 100 pl picoliter for an array of 49 (7×7) mini-microspots per microtitre wells (total coated antibody area= 10^6 μm^2) for 16 wells. The wells were blocked with SuperBlock and washed with phosphate buffer as above. The coated antibody density for both micro and mini-microspots are estimated to be 2×10^4 IgG/μm^2.

3. 200 μl of plasma containing 1 μU/ml of TSH was added to all the microtitre wells and shaken at room temperature. At 30, 60, 120 min and 18 hours (overnight), four wells containing the microspots and mini-microspots were washed with phosphate buffer containing 0.1% TWEEN 20, then incubated with 200 μl of anti-TSH developing antibody conjugated to hydrophilic latex microspheres in Tris-HCl buffer (50 μg/ml) for 30 min at room temperature and washed with phosphate-TWEEN 20 buffer. The wells were then scanned with a laser scanning confocal microscope equiped with an Argon/Krypton laser.

Results

Sample incubation	Total Fluorescent Signal (arbitrary units)	
times (mins)	Microspot	Mini-microspot
30	85 ± 7	111 ± 13
60	118 ± 15	149 ± 16
120	141 ± 21	178 ± 16
Overnight	185 ± 20	191 ± 23

Conclusion

Significantly higher mean responses were observed between 30 and 120 mins in the mini-micro spot samples, while the overnight controls did not show significant differences. This demonstrates that the mini-microspots have faster kinetic for the association of analyte with the capture antibody, and could be used to reduce incubation times.

The invention claimed is:

1. A method for determining the concentration of at least one analyte in a liquid sample, said method comprising, for each analyte, the steps of:

(a) immobilizing a specific binding agent including binding sites specific for the analyte on a solid support, wherein the specific binding agent used to determine the concentration of the analyte is present in an amount less than 0.1 V/K moles, where V is the volume of the liquid sample and K is the association constant for the analyte specifically binding to the specific binding agent, and wherein said specific binding agent is divided into an array of spatially separated locations;

(b) contacting the support with the sample so that a fraction of the binding sites of the specific binding agent specific for the analyte specifically binds the analyte;

(c) contacting the support with a developing agent labelled with a signal-producing marker such that the labelled developing agent binds to unoccupied binding sites, to specifically bound analyte or to the binding sites with specifically bound analyte;

(d) separating non-specifically bound developing agent from the solid support and measuring the signal produced by the marker at each of the locations in the array to obtain a value which represents the fraction of the binding sites occupied by the analyte at each location;

(e) adding the values obtained at the locations in the array to provide a total signal; and

(f) comparing the total signal to corresponding values obtained from a series of standard solutions containing known concentrations of the analyte, to determine the concentration of the analyte in the liquid sample.

2. The method according to claim 1, wherein the specific binding agent is divided into between 4 and 40 locations.

3. The method according to claim 1, wherein the locations have an area of about 10000 μm^2, the locations being separated from each other by a distance of 100 to 1000 μm.

4. The method according to claim 1, wherein the concentration of a plurality of different analytes in the liquid sample are determined using a plurality of arrays on said support.

5. The method according to claim 1, wherein (i) the specific binding agent is an antibody and the analyte is an antigen or (ii) the specific binding agent is an oligonucleotide and the analyte is a nucleic acid.

6. A method for determining the concentration of at least one analyte in a liquid sample, said method comprising, for each analyte, the steps of:

(a) immobilizing a specific binding agent including binding sites specific for the analyte on a solid support, wherein the specific binding agent used to determine the concentration of the analyte is present in an amount less than 0.1 V/K moles, where V is the volume of the liquid sample and K is the association constant for the analyte specifically binding to the specific binding agent, and wherein said specific binding agent is divided into an array of spatially separated locations;

(b) contacting the support with the liquid sample so that a fraction of the binding sites of the binding agent specific for the analyte specifically bind the analyte;

(c) contacting the support with a developing agent labelled with a signal-producing marker such that the labelled developing agent binds to unoccupied binding sites, to specifically bound analyte or to the binding sites with specifically bound analyte;

(d) separating non-specifically bound developing agent from the solid support and measuring the signal produced by the marker at each of the locations in the array to obtain a value which represents the fraction of the binding sites occupied by the analyte at each location; and

(e) adding the values obtained at the locations in the array to provide a total signal which indicates the concentration of the analyte in the liquid sample.

7. The method according to claim 6 wherein the specific binding agent is divided into between 4 and 40 locations.

8. The method according to claim 6, wherein the locations are in an area of about 10000 μm^2, the locations being separated from each other by a distance of 100 to 1000 μm.

9. The method according to claim 6, wherein the concentrations of a plurality of different analytes in the liquid sample are determined using a plurality of arrays on said support.

10. The method according to claim 6, wherein (i) the specific binding agent is an antibody and the analyte is an antigen or (ii) the specific binding agent is an oligonucleotide and the analyte is a nucleic acid.

11. A method for determining the concentration of at least one analyte in a liquid sample, said method employing a solid support on which is immobilized, for each analyte, a specific binding agent including binding sites specific for the analyte, wherein the specific binding agent used to determine the concentration of the analyte is present in an amount less than 0.1 V/K moles, where V is the volume of the liquid sample and K is the association constant for the analyte specifically binding to the specific binding agent, and wherein said specific binding agent is divided into an array of spatially separated locations, said method comprising the steps of:

(a) contacting the support with the sample so that a fraction of the binding sites of the specific binding agent specific for the analyte specifically binds the analyte;

(b) contacting the support with a developing agent labelled with a signal-producing marker such that the labelled developing agent binds to unoccupied binding sites, to specifically bound analyte or to the binding sites with specifically bound analyte;

(c) separating non-specifically bound developing agent from the solid support and measuring the signal produced by the marker at each of the locations in the array to obtain a value which represents the fraction of the binding sites occupied by the analyte at each location;

(d) adding the values obtained at the locations in the array to provide a total signal; and

(e) comparing the total signal to corresponding values obtained from a series of standard solutions containing known concentrations of the analyte, to determine the concentration of the analyte in the liquid sample.

12. The method according to claim 11, wherein the specific binding agent is divided into between 4 and 40 locations.

13. The method according to claim 11, wherein the locations have an area of about 10000 μm^2, the locations being separated from each other by a distance of 100 to 1000 μm.

14. The method according to claim 11, wherein the concentrations of a plurality of different analytes in the liquid sample are determined using a plurality of arrays on said support.

15. The method according to claim 11, wherein the specific binding agent is an antibody and the analyte is an antigen.

16. The method according to claim 11, wherein the specific binding agent is an oligonucleotide and the analyte is a nucleic acid.

17. A method for determining a value representative of a fraction of binding sites of a specific binding agent including binding sites specific for an analyte which binding sites are occupied by the analyte present in a liquid sample, said method comprising the steps of:

(a) immobilizing the specific binding agent on a solid support, wherein the specific binding agent used for the fractional occupancy determination is present in an amount less than 0.1 V/K moles, where V is the volume of the liquid sample and K is the association constant for the analyte specifically binding to the specific binding agent, and wherein said specific binding agent is divided into an array of spatially separated locations;

(b) contacting the support with the liquid sample so that a fraction of the binding sites of the binding agent specific for the analyte specifically bind the analyte;

(c) contacting the support with a developing agent labelled with a signal-producing marker such that the labelled developing agent binds to unoccupied binding sites, to specifically bound analyte or to the binding sites with specifically bound analyte;

(d) separating non-specifically bound developing agent from the solid support and measuring the signal produced by the marker at each of the locations in the array to obtain a value which represents the fraction of the binding sites occupied by the analyte at each location; and

(e) adding the values obtained at the locations in the array to provide a total signal which indicates the fraction of the binding sites in the specific binding agent occupied by the analyte.

18. The method according to claim 17, wherein the specific binding agent is divided into between 4 and 40 locations.

19. The method according to claim 17, wherein the locations are in an area of about 10000 μm^2, the locations being separated from each other by a distance of 100 to 1000 μm.

20. The method according to claim 17, wherein the fraction of occupied binding sites is determined for a plurality of different analytes in the liquid sample using a plurality of arrays on said support.

21. The method according to claim 17, wherein (i) the specific binding agent is an antibody and the analyte is an antigen or (ii) the specific binding agent is an oligonucleotide and the analyte is a nucleic acid.

* * * * *

US005569588A

United States Patent [19]

Ashby et al.

[11] **Patent Number: 5,569,588**

[45] **Date of Patent: Oct. 29, 1996**

[54] **METHODS FOR DRUG SCREENING**

[75] Inventors: **Matthew Ashby**, San Aselmo; **Jasper Rine**, Moraga, both of Calif.

[73] Assignee: **The Regents of the University of California**, Oakland, Calif.

[21] Appl. No.: **512,811**

[22] Filed: **Aug. 9, 1995**

[51] **Int. Cl.**^{6} **C12Q 1/68**; C12N 15/00; C07H 21/04

[52] **U.S. Cl.** 435/6; 435/29; 435/172.1; 536/23.4; 536/24.1

[58] **Field of Search** 435/6, 172.1, 172.3; 536/23.4, 24.1; 935/23, 47

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,981,784	1/1991	Evans et al.	435/6
5,378,603	1/1995	Brown et al.	435/6

FOREIGN PATENT DOCUMENTS

92304902.7	12/1992	European Pat. Off. .
WO92/05286	9/1990	WIPO .
WO94/17208	8/1994	WIPO .

Primary Examiner—George C. Elliott
Assistant Examiner—John S. Brusca
Attorney, Agent, or Firm—Richard Aron Osman, PH.D.

[57] **ABSTRACT**

Methods and compositions for modeling the transcriptional responsiveness of an organism to a candidate drug involve (a) detecting reporter gene product signals from each of a plurality of different, separately isolated cells of a target organism, wherein each cell contains a recombinant construct comprising a reporter gene operatively linked to a different endogenous transcriptional regulatory element of the target organism such that the transcriptional regulatory element regulates the expression of the reporter gene, and the sum of the cells comprises an ensemble of the transcriptional regulatory elements of the organism sufficient to model the transcriptional responsiveness of said organism to a drug; (b) contacting each cell with a candidate drug; (c) detecting reporter gene product signals from each cell; (d) comparing reporter gene product signals from each cell before and after contacting the cell with the candidate drug to obtain a drug response profile which provides a model of the transcriptional responsiveness of said organism to the candidate drug.

8 Claims, No Drawings

METHODS FOR DRUG SCREENING

BACKGROUND

The field of the invention is pharmaceutical drug screening. Pharmaceutical research and development is a multi-billion dollar industry. Much of these resources are consumed in efforts to focus the specificity of lead compounds. In addition, many programs are aborted after decades of costly yet fruitless efforts to limit side effects or toxicity of candidate drugs. Accordingly, tools that can abbreviate the research and discovery phase of drug development are desirable. Several in vitro or cell culture-based methods have been described for identifying compounds with a particular biological effect through the activation of a linked reporter. Gadski et al. (1992) EP 92304902.7 describes methods for identifying substances which regulate the synthesis of an apolipoprotein; Evans et al. (1991) U.S. Pat. No. 4,981,784 describes methods for identifying ligand for a receptor and Farr et al. (1994) WO 94/17208 describes methods and kits utilizing stress promoters to determine toxicity of a compound.

In general, the principle that has been applied in the existing pharmaceutical industry for the discovery and development of new lead compounds for drugs has been the establishment of sensitive and reliable in vitro assays for purified enzymes, and then screening large numbers of compounds and culture supernatants for any ability to inhibit enzyme activity. The present invention exploits the recent advances in genome science to provide for the rapid screening of large numbers of compounds against a systemic target comprising substantially all targets in a pathway, organism, etc. for rare compounds having the ability to inhibit the protein of interest. The invention described herein, in effect, turns the drug discovery process inside out. This invention provides information on the mechanism of action of every compound that affects cells, regardless of the target. In addition, the relative specificity of all lead compounds is immediately established.

SUMMARY OF THE INVENTION

The invention provides methods and compositions for estimating the physiological specificity of a candidate drug. In general, the subject methods involve (a) detecting reporter gene product Signals from each of a plurality of different, separately isolated cells of a target organism, wherein each of said cells contains a recombinant construct comprising a reporter gene operatively linked to a different endogenous transcriptional regulatory element (e.g. promoter) of said target organism such that said transcriptional regulatory element regulates the expression of said reporter gene, wherein said plurality of cells comprises an ensemble of the transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug; (b) contacting each said cell with a candidate drug; (c) detecting reporter gene product signals from each of said cells; (d) comparing said reporter gene product signals from each of said cells before and after contacting each of said cells with said candidate drug to obtain a drug response profile; wherein said drug response profile provides an estimate of the physiological specificity or biological interactions of said candidate drug.

DETAILED DESCRIPTION OF THE INVENTION

The Genome Reporter Matrix.

The invention provides methods and compositions for estimating the physiological specificity of a candidate drug by modeling the transcriptional responses of the target organism with an ensemble of reporters, the expressions of which are regulated by transcription regulatory genetic elements derived from the genome of the target organism. The ensemble of reporting cells comprises as comprehensive a collection of transcription regulatory genetic elements as is conveniently available for the targeted organism so as to most accurately model the systemic transcriptional response. Suitable ensembles generally comprise thousands of individually reporting elements; preferred ensembles are substantially comprehensive, i.e. provide a transcriptional response diversity comparable to that of the target organism. Generally, a substantially comprehensive ensemble requires transcription regulatory genetic elements from at least a majority of the organism's genes, and preferably includes those of all or nearly all of the genes. We term such a substantially comprehensive ensemble a genome reporter matrix.

It is frequently convenient to use an ensemble or genome reporter matrix derived from a lower eukaryote or common animal model to obtain preliminary information on drug specificity in higher eukaryotes, such as humans. Because yeast, such as *Saccharomyces cerevisiae*, is a bona fide eukaryote, there is substantial conservation of biochemical function between yeast and human cells in most pathways, from the sterol biosynthetic pathway to the Ras oncogene. Indeed, the absence of many effective antifungal compounds illustrates how difficult it has been to find therapeutic targets that would selectively kill fungal but not human cells. One example of a shared response pathway is sterol biosynthesis. In human cells, the drug Mevacor (lovastatin) inhibits HMG-CoA reductase, the key regulatory enzyme of the sterol biosynthetic pathway. As a result, the level of a particular regulatory sterol decreases, and the cells respond by increased transcription of the gene encoding the LDL receptor. In yeast, Mevacor also inhibits HMG-CoA reductase and lowers the level of a key regulatory sterol. Yeast cells respond in an analogous fashion to human cells. However, yeast do not have a gene for the LDL receptor. Instead, the same effect is measured by increased transcription of the ERG 10 gene, which encodes acetoacetyl CoA thiolase, an enzyme also involved in sterol synthesis. Thus the regulatory response is conserved between yeast and humans, even though the identity of the responding gene is different.

Advantages of the Genome Reporter Matrix as a Vehicle for Pharmaceutical Development

The advantages of the subject methods over prior art screening methods may be illustrated by examples. Consider the difference between an in vitro assay for HMG-CoA reductase inhibitors as presently practiced by the pharmaceutical industry, and an assay for inhibitors of sterol biosynthesis as revealed by the ERG 10 reporter. In the case of the former, information is obtained only for those rare compounds that happen to inhibit this one enzyme. In contrast, in the case of the ERG 10 reporter, any compound that inhibits nearly any of the approximately 35 steps in the sterol biosynthetic pathway will, by lowering the level of intracellular sterols, induce the synthesis of the reporter. Thus, the reporter can detect a much broader range of targets than can the purified enzyme, in this case 35 times more than the in vitro assay.

Drugs often have side effects that are in part due to the lack of target specificity. However, the in vitro assay of HMG-CoA reductase provides no information on the specificity of a compound. In contrast, a genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reporters, the first compound is, a priori, more likely to have fewer side effects. Because the identity of the reporters is known or determinable, information on other affected reporters is informative as to the nature of the side effect. A panel of reporters can be used to test derivatives of the lead compound to determine which of the derivatives have greater specificity than the first compound.

As another example, consider the case of a compound that does not affect the in vitro assay for HMG-CoA reductase nor induces the expression of the ERG10 reporter. In the traditional approach to drug discovery, a compound that does not inhibit the target being tested provides no useful information. However, a compound having any significant effect on a biological process generally has some consequence on gene expression. A genome reporter matrix can thus provide two different kinds of information for most compounds. In some cases, the identity of reporter genes affected by the inhibitor evidences to how the inhibitor functions. For example, a compound that induces a cAMP-dependent promoter in yeast may affect the activity of the Ras pathway. Even where the compound affects the expression of a set of genes that do not evidence the action of the compound, the matrix provides a comprehensive assessment of the action of the compound that can be stored in a database for later analyses. A library of such matrix response profiles can be continuously investigated, much as the Spectral Compendiums of chemistry are continually referenced in the chemical arts. For example, if the database reveals that compound X alters the expression of gene Y, and a paper is published reporting that the expression of gene Y is sensitive to, for example, the inositol phosphate signaling pathway, compound X is a candidate for modulating the inositol phosphate signaling pathway. In effect the genome reporter matrix is an informational translator that takes information on a gene directly to a compound that may already have been found to affect the expression of that gene. This tool should dramatically shorten the research and discovery phase of drug development, and effectively leverage the value of the publicly available research portfolio on all genes.

In many cases, a drug of interest would work on protein targets whose impact on gene expression would not be known a priori. The genome reporter matrix can nevertheless be used to estimate which genes would be induced or repressed by the drug. In one embodiment, a dominant mutant form of the gene encoding a drug-targeted protein is introduced into all the strains of the genome reporter matrix and the effect of the dominant mutant, which interferes with the gene product's normal function, evaluated for each reporter. This genetic assay informs us which genes would be affected by a drug that has a similar mechanism of action. In many cases, the drug itself could be used to obtain the same information. However, even if the drug itself were not available, genetics can be used to predetermine what its response profile would be in the genome reporter matrix. Furthermore, it is not necessary to know the identity of any of the responding genes. Instead, the genetic control with the dominant mutant sorts the genome into those genes that respond and those that do not. Hence, if drugs that disrupt a given cellular function were desired, dominant mutants for such function introduced into the genome reporter matrix reveal what response profile to expect for such an agent.

For example, taxol, a recent advance in potential breast cancer therapies, has been shown to interfere with tubulin-based cytoskeletal elements. Hence, a dominant mutant form of tubulin provides a response profile informative for breast cancer therapies with similar modes of action to taxol. Specifically, a dominant mutant form of tubulin is introduced into all the strains of the genome reporter matrix and the effect of this dominant mutant, which interferes with the microtubule cytoskeleton, evaluated for each reporter. Thus, any new compound that induces the same response profile as the dominant tubulin mutant would provide a candidate for a taxol-like pharmaceutical.

In addition, the genome reporter matrix can be used to genetically create or model various disease states. In this way, pathways present specifically in the disease state can be targeted. For example, the specific response profile of transforming mutant Ras2^{val19} identifies Ras2^{val19} induced reporters. Here, the matrix, in which each unit contains the Ras2^{val19} mutation is used to screen for compounds that restore the response profile to that of the matrix lacking the mutation.

Though these examples are directed to the development of human therapeutics, informative response profiles can often be obtained in nonhuman reporter matrices. Hence, for disease causing genes with yeast homologs, even if the function of the gene is not known, a dominant form of the gene can be introduced into a yeast-based reporter matrix to identify disease state specific pathways for targeting. For example, a reporter matrix comprising the yeast mutant Ras2^{val19} provides a discovery vehicle for pathways specific to the human analog, the oncogene Ras2^{val12}.

Application of Novel Combinatorial Chemistries with the Genome Reporter Matrix.

Among the most important advances in drug development have been advances in combinatorial synthesis of chemical libraries. In conventional drug screening with purified enzyme targets, combinatorial chemistries can often help create new derivatives of a lead compound that will also inhibit the target enzyme but with some different and desirable property. However, conventional methods would fail to recognize a molecule having a substantially divergent specificity. The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters. The identity of that gene provides a guide to the target of the new compound. Furthermore, the matrix offers an added bonus that compensates for a common weakness in most chemical syntheses. Specifically, most syntheses produce the desired product in greatest abundance and a collection of other related products as contaminants due to side reactions in the synthesis. Traditionally the solution to contaminants is to purify away from them. However, the genome reporter matrix exploits the presence of these contaminants. Syntheses can be adjusted to make them less specific with a greater number of side reactions and more contaminants to determine whether anything in the total synthesis affects the expression of target genes of interest. If there is a component of the mixture with the desired activity on a particular reporter, that reporter can be used to assay purification of the desired component from the mixture. In effect, the reporter matrix allows a focused survey of the effect on single genes to compensate for the impurity of the mixture being tested.

Isoprenoids are a specially attractive class for the genome reporter matrix. In nature, isoprenoids are the champion signaling molecules. Isoprenoids are derivatives of the five carbon compound isoprene, which is made as an intermediate in cholesterol biosynthesis. Isoprenoids include many of the most famous fragrances, pigments, and other biologically active compounds, such as the antifungal sesquiterpenoids, which plants use defensively against fungal infection. There are roughly 10,000 characterized isoprene derivatives and many more potential ones. Because these compounds are used in nature to signal biological processes, they are likely to include some of the best membrane permeant molecules.

Isoprenes possess another characteristic that lends itself well to drug discovery through the genome reporter matrix. Pure isoprenoid compounds can be chemically treated to create a wide mixture of different compounds quickly and easily, due to the particular arrangement of double bonds in the hydrocarbon chains. In effect, isoprenoids can be mutagenized from one form into many different forms much as a wild-type gene can be mutagenized into many different mutants. For example, vitamin D used to fortify milk is produced by ultraviolet irradiation of the isoprene derivative known as ergosterol. New biologically active isoprenoids are generated and analyzed with a genome reporter matrix as follows. First a pure isoprenoid such as limonene is tested to determine its response profile across the matrix. Next, the isoprenoid (e.g. limonene) is chemically altered to create a mixture of different compounds. This mixture is then tested across the matrix. If any new responses are observed, then the mixture has new biologically active species. In addition the identity of the reporter genes provides information regarding what the new active species does, an activity to be used to monitor its purification, etc. This strategy is also applied to other mutable chemical families in addition to isoprenoids.

Applications of the Genome Reporter Matrix in Antibiotic and Antifungal Discovery.

Fungi are important pathogens on plants and animals and make a major impact on the production of many food crops and on animal, including human, health. One major difficulty in the development of antifungal compounds has been the problem of finding pharmaceutical targets in fungi that are specific to the fungus. The genome reporter matrix offers a new tool to solve this problem. Specifically, all molecules that fail to elicit any response in the Saccharomyces reporter are collected into a set, which by definition must be either inactive biologically or have a very high specificity. A reporter library is created from the targeted pathogen such as Cryptococcus, Candida, Aspergillus, Pneumocystis etc. All molecules from the set that do not affect Saccharomyces are tested on the pathogen, and any molecule that elicits an altered response profile in the pathogen in principle identifies a target that is pathogen-specific. As an example, a pathogen may have a novel signaling enzyme, such as an inositol kinase that alters a position on the inositol ring that is not altered in other species. A compound that inhibits that enzyme would affect the signaling pathway in the pathogen, and alter a response profile, but due to the absence of that enzyme in other organisms, would have no effect. By sequencing the reporter genes affected specifically in the target fungus and comparing the sequence with others in Genbank, one can identify biochemical pathways that are unique to the target species. Useful identified products include not only agents that kill the target fungus but also the identification of specific targets in the fungus for other pharmaceutical screening assays.

The identification of compounds that kill bacteria has been successfully pursued by the pharmaceutical industry for decades. It is rather simple to spot a compound that kills bacteria in a spot test on a petri plate. Unfortunately, growth inhibition screens have provided very limited lead compound diversity. However, there is much complexity to bacterial physiology and ecology that could offer an edge to development of combination therapies for bacteria, even for compounds that do not actually kill the bacterial cell. Consider for example the bacteria that invade the urethra and persist there through the elaboration of surface attachments known as timbrae. Antibiotics in the urine stream have limited access to the bacteria because the urine stream is short-lived and infrequent. However, if one could block the synthesis of the timbrae to detach the bacteria, existing therapies would become more effective. Similarly, if the chemotaxis mechanism of bacteria were crippled, the ability of bacteria to establish an effective infection would, in some species, be compromised. A genome reporter matrix for a bacterial pathogen that contains reporters for the expression of genes involved in chemotaxis or fimbrae synthesis, as examples, identifies not only compounds that do kill the bacteria in a spot test, but also those that interfere with key steps in the biology of the pathogen. These compounds would be exceedingly difficult to discover by conventional means.

Applications of Human Cell Based Genome Reporter Matrices.

A genome reporter matrix based on human cells provides many important applications. For example, an interesting application is the development of antiviral compounds. When human cells are infected by a wide range of viruses, the cells respond in a complex way in which only a few of the components have been identified. For example, certain interferons are induced as is a double-stranded RNase. Both of these responses individually provides some measure of protection. A matrix that reports the induction of interferon genes and the double stranded RNase is able to detect compounds that could prophylactically protect cells before the arrival of the virus. Other protective effects may be induced in parallel. The incorporation of a panel of other reporter genes in the matrix is used to identify those compounds with the highest degree of specificity.

Use of the Genome Reporter Matrix.

The procedure to be followed in the subject methods will now be outlined. The initial step involves determining the basal or background response profile by detecting reporter gene product signals from each of a plurality of different, separately isolated cells of a target organism under one or more of a variety of physical conditions, such as temperature and pH, medium, and osmolarity. As discussed above, the target organism may be a yeast, animal model, human, plant, pathogen, etc. Generally, the cells are arranged in a physical matrix such as a microtiter plate. Each of the cells contains a recombinant construct comprising a reporter gene operatively linked to a different endogenous transcriptional regulatory element of said target organism such that said transcriptional regulatory element regulates the expression of said reporter gene. A sufficient number of different recombinant cells are included to provide an ensemble of transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug. In a preferred embodiment, the matrix is substantially comprehensive for the selected regulatory elements, e.g. essentially all of the gene promoters of the targeted organism are included. Other cis-acting or trans-acting transcription regulatory regions of the targeted organism can also be evaluated. In one embodiment, a genome reporter matrix is constructed from a set of lacZ fusions to a substantially comprehensive set of yeast genes. The fusions are preferably constructed in a diploid cell of the a/a mating type to allow the introduction of dominant mutations by mating, though haploid strains also find use with particularly sensitive reporters for certain functions. The fusions are conveniently arrayed onto a microtiter plate having 96 wells separating distinct fusions into wells having defined alphanumeric X-Y coordinates, where each well (defined as a unit) confines a cell or colony of cells having a construct of a reporter gene operatively joined to a different transcriptional promoter. Permanent collections of these plates are readily maintained at −80° C. and copies of this collection can be made and propagated by simple mechanics and may be automated with commercial robotics.

The methods involve detecting a reporter gene product signal for each cell of the matrix. A wide variety of reporters may be used, with preferred reporters providing conveniently detectable signals (e.g. by spectroscopy). Typically, the signal is a change in one or more electromagnetic properties, particularly optical properties at the unit. As examples, a reporter gene may encode an enzyme which catalyzes a reaction at the unit which alters light absorption properties at the unit, radiolabeled or fluorescent tag-labeled nucleotides can be incorporated into nascent transcripts which are then identified when bound to oligonucleotide probes, etc. Examples include β-galactosidase, invertase, green fluorescent protein, etc. Invertase fusions have the virtue that functional fusions can be selected from complex libraries by the ability of invertase to allow those genes whose expression increases or decreases by measuring the relative growth on medium containing sucrose with or without the compound of interest. Electronic detectors for optical, radiative, etc. signals are commercially available, e.g. automated, multi-well colorimetric detectors, similar to automated ELISA readers. Reporter gene product signals may also be monitored as a function of other variables such as stimulus intensity or duration, time (for dynamic response analyses), etc.

In a preferred embodiment, the basal response profiles are determined through the colorimetric detection of a lacZ reaction product. The optical signal generated at each well is detected and linearly transduced to generate a corresponding digital electrical output signal. The resultant electrical output signals are stored in computer memory as a genome reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus or drug. This information is indexed against the matrix to form reference response profiles that are used to determine the response of each reporter to any milieu in which a stimulus may be provided.

After establishing a basal response profile for the matrix, each cell is contacted with a candidate drug. The term drug is used loosely to refer to agents which can provoke a specific cellular response. Preferred drugs are pharmaceutical agents, particularly therapeutic agents. The drug induces a complex response pattern of repression, silence and induction across the matrix (i.e. a decrease in reporter activity at some units, an increase at others, and no change at still others). The response profile reflects the cell's transcriptional adjustments to maintain homeostasis in the presence of the drug. While a wide variety of candidate drugs can be evaluated, it is important to adjust the incubation conditions (e.g. concentration, time, etc.) to preclude cellular stress, and hence insure the measurements of pharmaceutically relevant response profiles. Hence, the methods monitor transcriptional changes which the cell uses to maintain cellular homeostasis. Cellular stress may be monitored by any convenient way such as membrane potential (e.g. dye exclusion), cellular morphology, expression of stress response genes, etc. In a preferred embodiment, the compound treatment is performed by transferring a copy of the entire matrix to fresh medium containing the first compound of interest.

After contacting the cells with the candidate drug, the reporter gene product signals from each of said cells is again measured to determine a stimulated response profile. The basal of background response profile is then compared with (e.g. subtracted from, or divided into) the stimulated response profile to identify the cellular response profile to the candidate drug. The cellular response can be characterized in a number of ways. For example, the basal profile can be subtracted from the stimulated profile to yield a net stimulation profile. In another embodiment, the stimulated profile is divided by the basal profile to yield an induction ratio profile. Such comparison profiles provide an estimate of the physiological specificity of the candidate drug.

In another embodiment of the invention, a matrix of hybridization probes corresponding to a predetermined population of genes of the selected organism is used to specifically detect changes in gene transcription which result from exposing the selected organism or cells thereof to a candidate drug. In this embodiment, one or more cells derived from the organism is exposed to the candidate drug in vivo or ex vivo under conditions wherein the drug effects a change in gene transcription in the cell to maintain homeostasis. Thereafter, the gene transcripts, primarily mRNA, of the cell or cells is isolated by conventional means. The isolated transcripts or cDNAs complementary thereto are then contacted with an ordered matrix of hybridization probes, each probe being specific for a different one of the transcripts, under conditions wherein each of the transcripts hybridizes with a corresponding one of the probes to form hybridization pairs. The ordered matrix of probes provides, in aggregate, complements for an ensemble of genes of the organism sufficient to model the transcriptional responsiveness of the organism to a drug. The probes are generally immobilized and arrayed onto a solid substrate such as a microtiter plate. Specific hybridization may be effected, for example, by washing the hybridized matrix with excess non-specific oligonucleotides. A hybridization signal is then detected at each hybridization pair to obtain a matrix-wide signal profile. A wide variety of hybridization signals may be used; conveniently, the cells are pre-labeled with radionucleotides such that the gene transcripts provide a radioactive signal that can be detected in the hybridization pairs. The matrix-wide signal profile of the drug-stimulated cells is then compared with a matrix-wide signal profile of negative control cells to obtain a specific drug response profile.

The invention also provides means for computer-based qualitative analysis of candidate drugs and unknown compounds. A wide variety of reference response profiles may be generated and used in such analyses. For example, the response of a matrix to loss of function of each protein or gene or RNA in the cell is evaluated by introducing a dominant allele of a gene to each reporter cell, and determining the response of the reporter as a function of the mutation. For this purpose, dominant mutations are preferred but other types of mutations can be used. Dominant mutations are created by in vitro mutagenesis of cloned genes followed by screening in diploid cells for dominant mutant alleles.

In an alternative embodiment, the reporter matrix is developed in a strain deficient for the UPF gene function, wherein the majority of nonsense mutations cause a dominant phenotype, allowing dominant mutations to be constructed for any gene. UPF1 encodes a protein that causes the degradation of MRNA's that, due to mutation, contain premature termination codons. In routants lacking UPF1 function most nonsense mutations encode short truncated protein fragments. Many of these interfere with normal protein function and hence have dominant phenotypes. Thus in a upf1 mutant, many nonsense alleles behave as dominant mutations (see, e.g. Leeds, P. et al. (1992) Molec. Cell Biology. 12:2165–77).

The resultant data identify genetic response profiles. These data are sorted by individual gene response to determine the specificity of each gene to a particular stimulus. A weighting matrix is established which weights the signals proportionally to the specificity of the corresponding reporters. The weighting matrix is revised dynamically, incorporating data from every screen. A gene regulation function is then used to construct tables of regulation identifying which cells of the matrix respond to which mutation in an indexed gene, and which mutations affect which cells of the matrix.

Response profiles for an unknown stimulus (e.g. new chemicals, unknown compounds or unknown mixtures) may be analyzed by comparing the new stimulus response profiles with response profiles to known chemical stimuli. Such comparison analyses generally take the form of an indexed report of the matches to the reference chemical response profiles, ranked according to the weighted value of each matching reporter. If there is a match (i.e. perfect score), the response profile identifies a stimulus with the same target as one of the known compounds upon which the response profile database is built. If the response profile is a subset of cells in the matrix stimulated by a known compound, the new compound is a candidate for a molecule with greater specificity than the reference compound. In particular, if the reporters responding uniquely to the reference chemical have a low weighted response value, the new compound is concluded to be of greater specificity. Alternatively, if the reporters responding uniquely to the reference compound have a high weighted response value, the new compound is concluded to be active downstream in the same pathway. If the output overlaps the response profile of a known reference compound, the overlap is sorted by a quantitative evaluation with the weighting matrix to yield common and unique reporters. The unique reporters are then sorted against the regulation tables and best matches used to deduce the candidate target. If the response profile does not either overlap or match a chemical response profile, then the database is inadequate to infer function and the response profile may be added to the reference chemical response profiles.

The response profile of a new chemical stimulus may also be compared to a known genetic response profile for target gene(s). If there is a match between the two response profiles, the target gene or its functional pathway is the presumptive target of the chemical. If the chemical response profile is a subset of a genetic response profile, the target of the drug is downstream of the mutant gene but in the same pathway. If the chemical response profile includes as a subset a genetic response profile, the target of the chemical is deduced to be in the same pathway as the target gene but upstream and/or the chemical affects additional cellular components. If not, the chemical response profile is novel and defines an orphan pathway.

While described in terms of cells comprising reporters under the transcriptional control of endogenous regulatory regions, there are a number of other means of practicing the invention. For example, each unit of a genome reporter matrix reporting on gene expression might confine a different oligonucleotide probe capable of hybridizing with a corresponding different reporter transcript. Alternatively, each unit of a matrix reporting on DNA-protein interaction might confine a cell having a first construct of a reporter gene operatively joined to a targeted transcription factor binding site and a second hybrid construct encoding a transcription activation domain fused to a different structural gene, i.e. a one-dimensional one-hybrid system matrix. Alternatively, each unit of a matrix reporting on protein-protein interactions might confine a cell having a first construct of a reporter gene operatively joined to a targeted transcription factor binding site, a second hybrid construct encoding a transcription activation domain fused to a different constitutionally expressed gene and a third construct encoding a DNA-binding domain fused to yet a different constitutionally expressed gene, i.e. a two-dimensional two-hybrid system matrix.

The following examples are offered by way of illustration and not by way of limitation.

EXAMPLES

1. Transcriptional promoter-reporter gene matrix

A) Construction of a physical matrix stimulated with the drug mevinolin (lovastatin, Meracon).

Mevinolin is a compound known to inhibit cholesterol biosynthesis. Initially, the maximal non-toxic (as measured by cell growth and viability) concentration of mevinolin on the reporter cells was determined by serial dilution to be 25 ug/ml. To produce a mevinolin-stimulated matrix, each well of 60 microtiter plates is filled with 100 ul culture medium containing 25 ug/ml mevinolin in a 2% ethanol solution. An aliquot of each member of the reporter matrix is added to each well allowing for a dilution of approximately 1:100. The cells are incubated in the medium until the turbidity of the average reporter increases by 20 fold. Each well is then quantified for turbidity as a measure of growth, and is treated with a lysis solution to allow measurement of β-galactosidase from each fusion.

B) Generation of an output signal matrix data structure.

Both the turbidity and the B-galactosidase are read on commercially available microtiter plate readers (e.g. Bio-Rad) and the data captured as an ASCII file. From this file, the value of the individual cells in the reporter matrix to a 2% ethanol solution in the reference response profile is subtracted. The difference corresponds to the mevinolin response profile. This file is converted in the computer to a table indexed by the response of each cell to the inhibitor. For example, the genes encoding acetoacetyl-CoA thiolase and squalene synthase increase 10 fold, while SIR3, and LEU2, two unrelated genes, remain unchanged. The response of the reporter matrix to other compounds is similarly determined and stored as output response profiles.

C) Comparison of Signal Matrix data structure with a Signal Matrix database.

A physical matrix is constructed as describe above except the mevinolin is replaced with an unknown test compound. The resultant response profile is compared to the response profiles of a library of known bioactive compounds and analyzed as described above. For example, if the test compound output profile shows both acetoacetyl-CoA thiolase and squalene synthase gene induced, then the output profile matches that expected of an inhibitor of cholesterol synthesis. If the response profile has fewer other cells affected than the response profile to mevinolin, the unknown compound is a candidate for greater specificity. If the response profile of the new chemical affects fewer other reporters than the response profile to mevinolin, and if the other reporters affected by mevinolin have a lower weighted value, then the compound is a candidate for greater specificity. If the response profile has more different cells affected than the response profile to mevinolin, then the compound is a candidate for less specificity. In the case where mixtures of compounds are tested, the highest weighted responses are evaluated to determine whether they can be deconvoluted into the response profile of two different compounds, or of two different genetic response profiles.

2. Reporter transcript-oligonucleotide hybridization probe matrix: Construction of stimulated physical matrix and generation of an output signal matrix data structure.

Unlabeled oligonucleotide hybridization probes complementary to the mRNA transcript of each yeast gene are arrayed on a silicon substrate etched by standard techniques (e.g. Fodor et al. (1991) Science 252, 767). The probes are of length and sequence to ensure specificity for the corresponding yeast gene, typically about 24–240 nucleotides in length.

A confluent HeLa cell culture is treated with 15 ug/ml mevinolin in 2% ethanol for 4 hours while maintained in a humidified 5% CO_2 atmosphere at 37° C. Messenger RNA is extracted, reverse transcribed and fluorophore-labeled according to standard methods (Sambrook et al., Molecular Cloning, 3rd ed.). The resultant cDNA is hybridized to the array of probes, the array is washed free of unhybridized labeled cDNA, the hybridization signal at each unit of the array quantified using a confocal microscope scanner (instruments by Molecular Devices and Affymetrix), and the resultant matrix response data stored in digital form.

3. Two-dimensional two-hybrid matrix

A) Construction of stimulated physical matrix.

The two-dimensional two-hybrid (see, e.g. Chien et al. (1991) PNAS, 88, 9578)matrix is designed to screen for compounds that specifically affect the interaction of two proteins, e.g. the interaction of a human signal transducer and activator of transcription (STAT) with an interleukin receptor. Two hybrid fusions are generated by standard methods: each strain contains a portion of the targeted human STAT gene, fused to a portion of a yeast or bacterial gene encoding a DNA binding domain (e.g. GAL4:1–147). The DNA sequence recognized by that DNA binding domain (e.g. UAS_G) is inserted in place of the enhancer sequence 5' to the selected reporter (e.g. lacZ). The strain also contains another fusion consisting of an intracellular portion of the targeted receptor gene whose protein product interacts with the STAT. This receptor gene is fused with a gene fragment encoding a transcriptional activation domain (e.g. GAL4:768–881).

B) Generation of signal matrix data structure.

Both the turbidity and the galactosidase are read on commercial microtiter plate readers (BioRad) and the data captured as an ASCII file.

C) Comparison of signal matrix data structure with database.

Data are analyzed for those compounds that block the interaction of the two human proteins by reducing the signal produced from the reporter in the various strains containing pairs of human proteins. The output is processed to identify compounds with a large impact on a reporter whose expression is dependent on a single pair of interacting human proteins. An inverted weighting matrix is used to evaluate these data as preferred compounds do not affect even the least specific reporters in the matrix.

All publications and patent applications cited in this specification are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

What is claimed is:

1. A method for modeling of the transcriptional responsiveness of an organism to a candidate drug which has an effect on gene transcription in cells of said organism, comprising steps:

(a) detecting reporter gene product signals from each of a plurality of different, separately isolated cells of a target organism, wherein each of said cells contains a recombinant construct comprising a reporter gene operatively linked to a different endogenous transcriptional regulatory element of said target organism such that said transcriptional regulatory element regulates the expression of said reporter gene, wherein said plurality of cells comprises an ensemble of the transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug;

(b) contacting each of said cells with a candidate drug under conditions wherein said cells maintain homeostasis;

(c) detecting reporter gene product signals from each of said cells;

(d) comparing said reporter gene product signals from each of said cells before and after contacting each of said cells with said candidate drug to obtain a drug response profile;

wherein said drug response profile provides a model of the transcriptional responsiveness of said organism to said candidate drug.

2. A method according to claim 1, said ensemble comprising a majority of all different transcriptional regulatory elements of said organism.

3. A method according to claim 1, said drug being a candidate human therapeutic.

4. A method according to claim 1, wherein said cells are yeast cells.

5. A method according to claim 1, wherein said cells are bacterial cells.

6. A method according to claim 1, wherein said cells are human cells.

7. A method according to claim 1, wherein the reporter gene is the lacZ gene, the suc2 gene, or a gene encoding a green fluorescent protein.

8. A method according to claim 1, wherein said cells are eukaryotic cells.

* * * * *

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ~0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

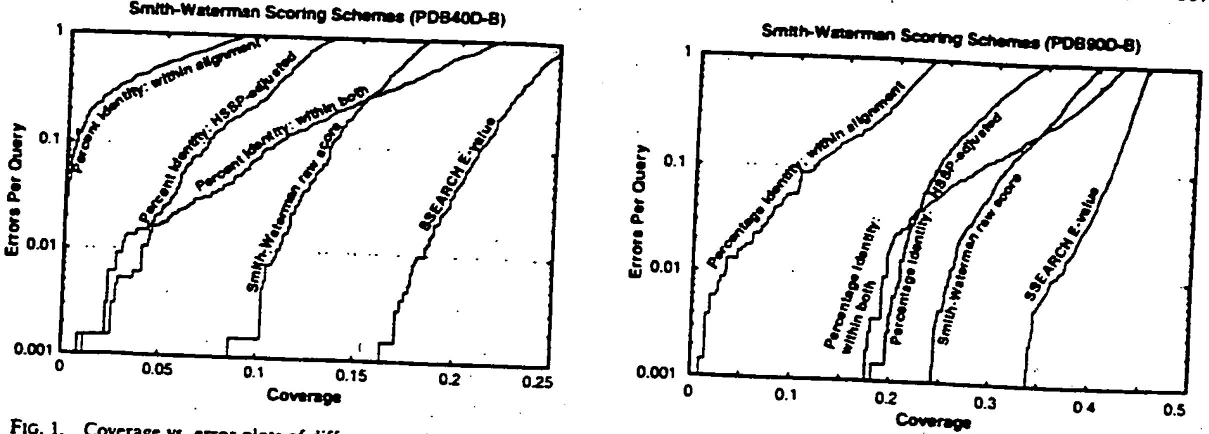


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the *x* axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The *y* axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The *y* axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

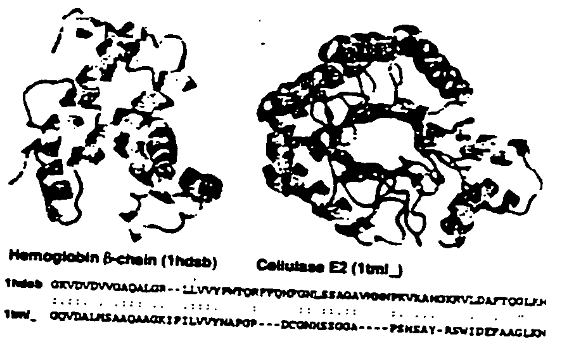


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

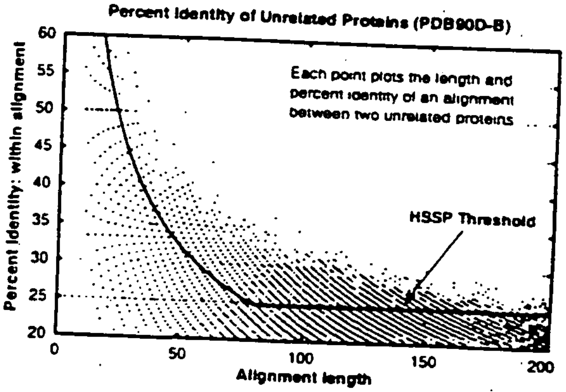


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

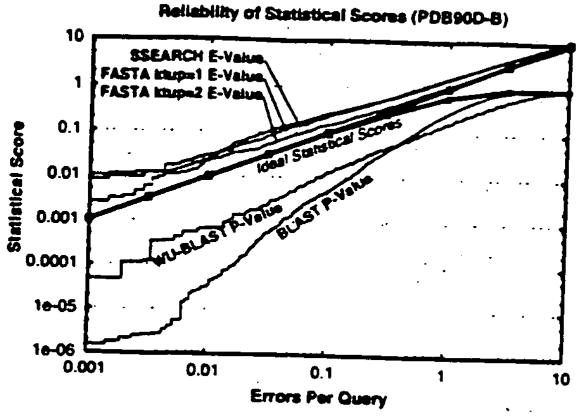


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

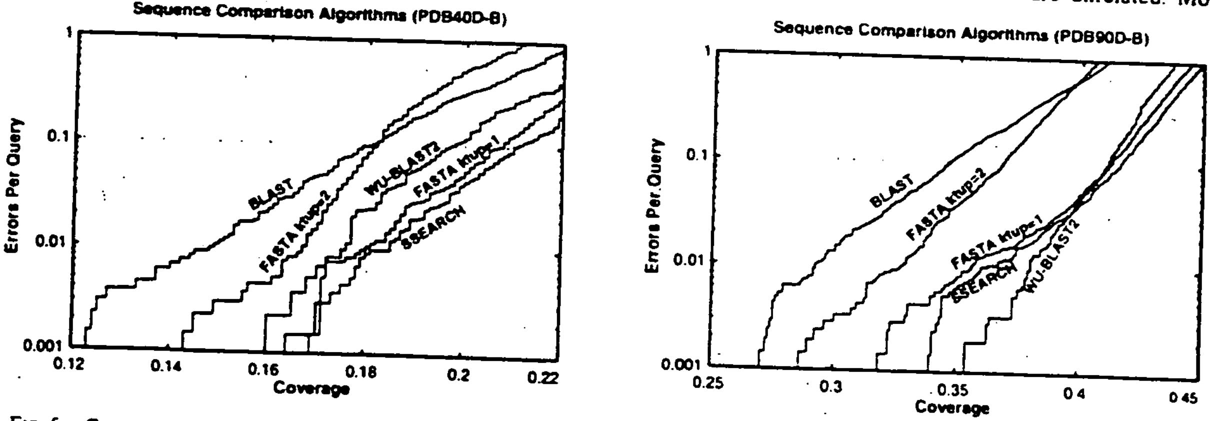


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure: it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

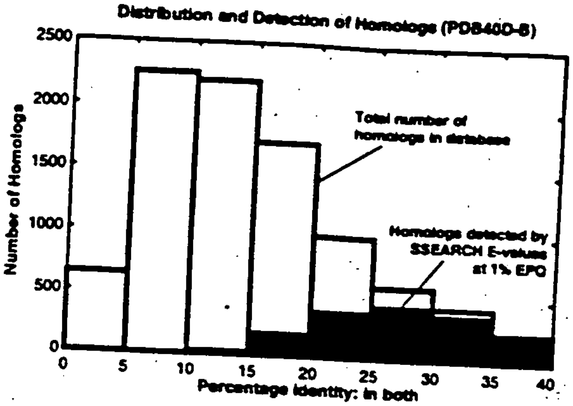


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* 247, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* 266, 635–643.
6. Pearson, W. R. (1991) *Genomics* 11, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* 4, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* 266, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* 249, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* 17, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* 24, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* 24, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* 9, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* 233, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* 1, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* 1, 77–78.
21. Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* 14, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* 90, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* 6, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* 266, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* 12, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* 266, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* 9, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* 13, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* 7, 369–376.
32. Orengo, C., Michie, A., Jones S. Jones D. T. Swindells M. B. & Thornton, J. (1997) *Structure (London)* 5, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* 39, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* 20, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* 16, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* 4, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* 131, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* 32, 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374–376.

Reference 7 of 11
of Response dated 12/04/03
In USSN 09/937,060

Subject: RE: [Fwd: T xicol gy Chip]
Date: Mon. 3 Jul 2000 08:09:45 -0400
From: "Afshari.Cynthia" <afshari@niehs.nih.gov>
To: ""Diana Hamlet-Cox"" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at
http: manuel.niehs.nih.gov maps guest clonesrch.cfm
W selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80-) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.
Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> *From:* *Diana Hamlet-Cox*
> *Sent:* *Monday, June 26, 2000 8:52 PM*
> *To:* *afshari@niehs.nih.gov*
> *Subject:* *[Fwd: Toxicology Chip]*
>
> *Dear Dr. Afshari,*
>
> *Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.*
>
> *Can you help me in this matter? I don't need to know the sequences, necessarily, but I would like very much to know what types of sequences are being used, e.g., GPCRs (more specific?), ion channels, etc.*
>
> *Diana Hamlet-Cox*
>
> *-------- Original Message --------*
> *Subject: Toxicology Chip*
> *Date: Mon, 19 Jun 2000 18:31:48 -0700*
> *From: Diana Hamlet-Cox <dianahc@incyte.com>*
> *Organization: Incyte Pharmaceuticals*
> *To: grigg@niehs.nih.gov*
>
> *Dear Colleague:*
>
> *I am doing literature research on the use of expressed genes as pharmacotoxicology markers, and found the Press Release dated February 29, 2000 regarding the work of the NIEHS in this area. I would like to know if there is a resource I can access (or you could provide?) that would give me a list of the 12,000 genes that are on your Human ToxChip Microarray. In particular, I am interested in the criteria used to select sequences for the ToxChip, including any control sequences included in the microarray.*
>
> *Thank you for your assistance in this request.*
>
> *Diana Hamlet-Cox, Ph.D.*
> *Incyte Genomics, Inc.*
>
> --
>
> ======================

> This email message is for the sole use of the intended recipient s and
> may contain confidential and privileged information subject to
> attorney-client privilege. Any unauthorized review, use, disclosure or
> distribution is prohibited. If you are not the intended recipient,
> please contact the sender by reply email and destroy all copies of the
> original message.
> ==========================
>
>

Reference 3 of 11
of Response dated 12/04/03
In USSN 09/937,060

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 8945–8947, August 1997
Applied Biological Sciences

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS*§

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT **The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.**

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5′ end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at http://www.pnas.org.

†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.

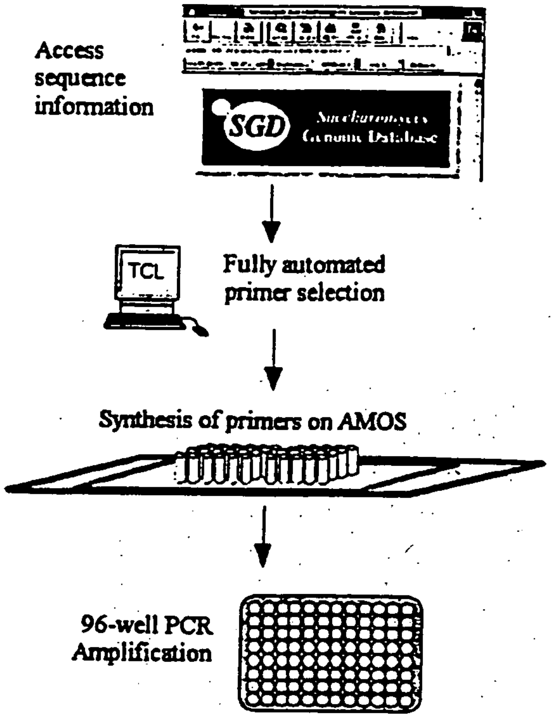

Access sequence information
SGD
Saccharomyces Genome Database
TCL
Fully automated primer selection
Synthesis of primers on AMOS
96-well PCR Amplification

FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose.

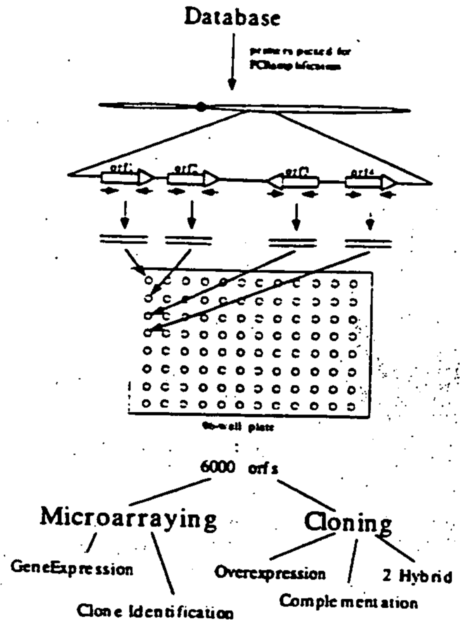


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

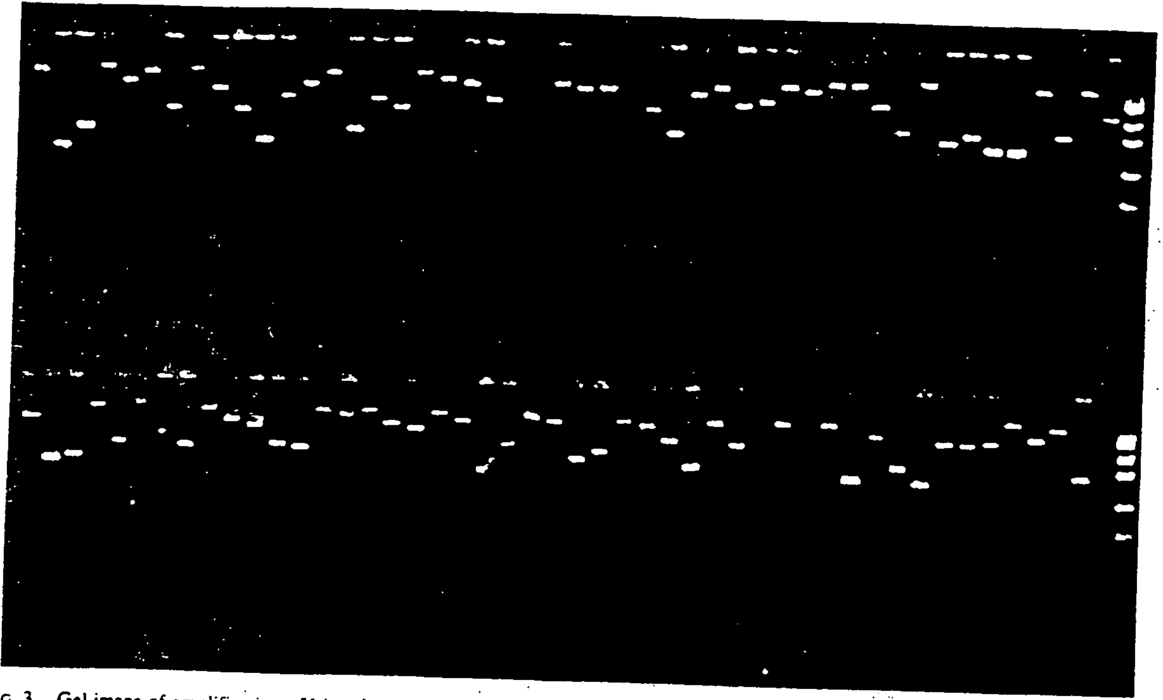


FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

Support was provided by National Institutes of Health Grants R37H60198 and P01H600205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269**, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270**, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273**, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356**, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106**, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379**, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57**, A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251**, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274**, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14**, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274**, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316**, 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* **340**, 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* **4**, 11–18.

Reference 2 of 11
of Response dated 12/04/03
In USSN 09/937,060

XENOBIOTICA , 1999, VOL. 29, NO. 7, 655–691

Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡
and G. GORDON GIBSON*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey, Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

Xenobiotica ISSN 0049-8254 print/ISSN 1366-5928 online © 1999 Taylor & Francis Ltd
http://www.tandf.co.uk/JNLS/xen.htm http://www.taylorandfrancis.com/JNLS/xen.htm

cell's development or response and should help in the elucidation of specific and sensitive biomarkers representing, for example, different types of cancer or previous exposure to certain classes of chemicals that are enzyme inducers.

In drug metabolism, many of the xenobiotic-metabolizing enzymes (including the well-characterized isoforms of cytochrome P450) are inducible by drugs and chemicals in man (Pelkonen *et al.* 1998), predominantly involving transcriptional activation of not only the cognate cytochrome P450 genes, but additional cellular proteins which may be crucial to the phenomenon of induction. Accordingly, the development of methodology to identify and assess the full complement of genes that are either up- or down-regulated by inducers are crucial in the development of knowledge to understand the precise molecular mechanisms of enzyme induction and how this relates to drug action. Similarly, in the field of chemical-induced toxicity, it is now becoming increasingly obvious that most adverse reactions to drugs and chemicals are the result of multiple gene regulation, some of which are causal and some of which are casually-related to the toxicological phenomenon *per se*. This observation has led to an upsurge in interest in gene-profiling technologies which differentiate between the control and toxin-treated gene pools in target tissues and is, therefore, of value in rationalizing the molecular mechanisms of xenobiotic-induced toxicity. Knowledge of toxin-dependent gene regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. For example, if the gene profile in response to say a testicular toxin that has been well-characterized *in vivo* could be determined in the testis, then this profile would be representative of all new drug candidates which act via this specific molecular mechanism of toxicity, thereby providing a useful and coherent approach to the early detection of such toxicants. Whereas it would be informative to know the identity and functionality of all genes up/down regulated by such toxicants, this would appear a longer term goal, as the majority of human genes have not yet been sequenced, far less their functionality determined. However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well-characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. Such approaches are beginning to gain momentum, in that several biotechnology companies are commercially producing 'gene chips' or 'gene arrays' that may be interrogated for toxicity assessment of xenobiotics. These chips consist of hundreds/thousands of genes, some of which are degenerate in the sense that not all of the genes are mechanistically-related to any one toxicological phenomenon. Whereas these chips are useful in broad-spectrum screening, they are maturing at a substantial rate, in that gene arrays are now becoming more specific, e.g. chips for the identification of changes in growth factor families that contribute to the aetiology and development of chemically-induced neoplasias.

Although documenting and explaining these genetic changes presents a formidable obstacle to understanding the different mechanisms of development and disease progression, the technology is now available to begin attempting this difficult challenge. Indeed, several 'differential expression analysis' methods have been developed which facilitate the identification of gene products that demonstrate altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinauer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

(1) Differential screening,
(2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
(3) Differential display (DD),
(4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
(5) Gene expression arrays, and
(6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter hybridization', which was used to isolate galactose-inducible DNA sequences from yeast. The theory is simple: a genomic DNA library is prepared from normal, unstimulated cells of the test organism/tissue and multiple filter replicas are prepared. These replica blots are probed with radioactively (or otherwise) labelled complex cDNA probes prepared from the control and test cell mRNA populations. Those mRNAs which are differentially expressed in the treated cell population will show a positive signal only on the filter probed with cDNA from the treated cells. Furthermore, labelled cDNA from different test conditions can be used to probe multiple blots, thereby enabling the identification of mRNAs which are only up-regulated under certain conditions. For example, St John and Davis (1979) screened replica filters with acetate-, glucose- and galactose-derived probes in order to obtain genes induced specifically by galactose metabolism. Although groundbreaking in its time this method is now considered insensitive and time-consuming, as up to 2 months are required to complete the identification of genes which are differentially expressed in the test population. In addition, there is no convenient way to check that the procedure has worked until the whole process has been completed.

Subtractive Hybridization (SH)

The developing concept of differential gene expression and the success of early approaches such as that described by St John and Davis (1979) soon gave rise to a search for more convenient methods of analysis. One of the first to be developed was SH, numerous variations of which have since been reported (see below). In general, this approach involves hybridization of mRNA/cDNA from one population (tester) to excess mRNA/cDNA from another (driver), followed by separation of the unhybridized tester fraction (differentially expressed) from the hybridized common sequences. This step has been achieved physically, chemically and through the use of selective polymerase chain reaction (PCR) techniques.

Physical separation

Original subtractive hybridization technology involved the physical separation of hybridized common species from unique single stranded species. Several methods of achieving this have been described, including hydroxyapatite chromatography (Sargent and Dawid 1983), avidin-biotin technology (Duguid and Dinauer 1990) and oligodT-latex separation (Hara *et al.* 1991). In the first approach, common mRNA species are removed by cDNA (from test cells)-mRNA (from control cells) subtractive hybridization followed by hydroxyapatite chromatography, as hydroxyapatite specifically adsorbs the cDNA-mRNA hybrids. The unabsorbed cDNA is then used either for the construction of a cDNA library of differentially expressed genes (Sargent and Dawid 1983, Schneider *et al.* 1988) or directly as a probe to screen a preselected library (Zimmerman *et al.* 1980, Davis *et al.* 1984, Hedrick *et al.* 1984). A schematic diagram of the procedure is shown in figure 1.

Less rigorous physical separation procedures coupled with sensitivity enhancing PCR steps were later developed as a means to overcome some of the problems encountered with the hydroxyapatite procedure. For example, Daguid and Dinauer (1990) described a method of subtraction utilizing biotin-affinity systems as a means to remove hybridized common sequences. In this process, both the control and tester mRNA populations are first converted to cDNA and an adaptor ('oligovector',

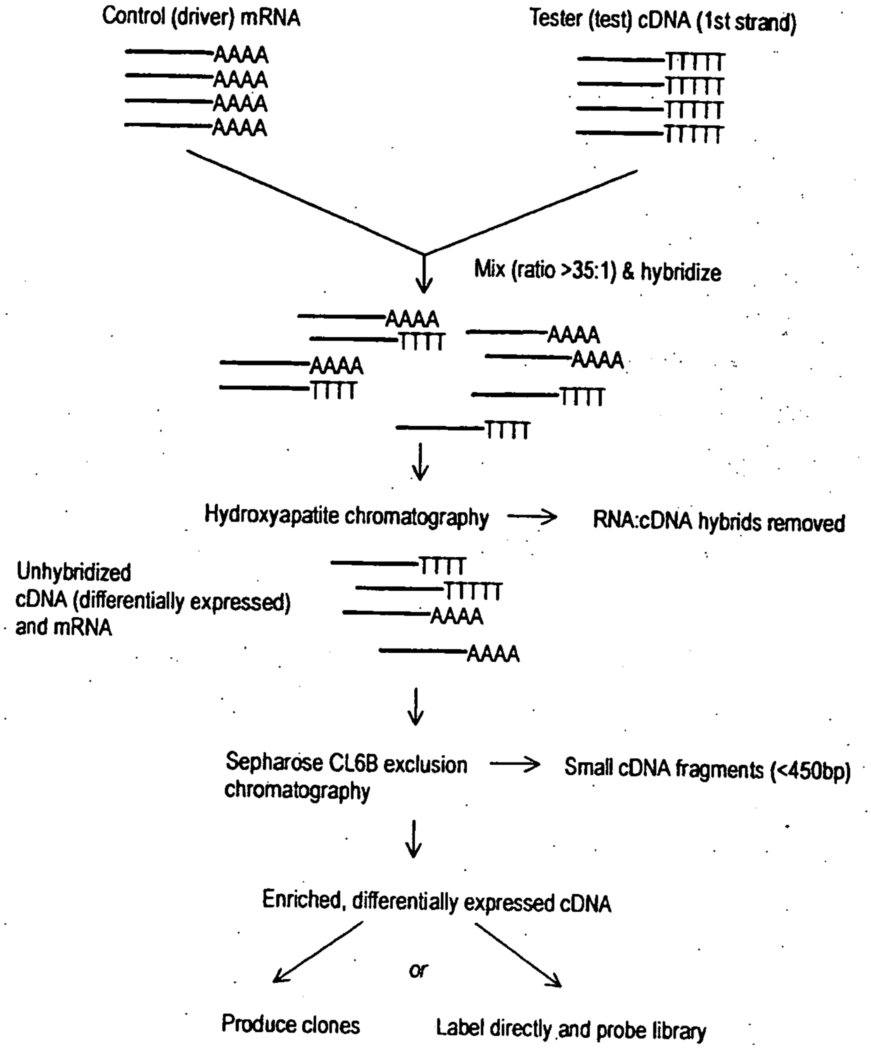

Control (driver) mRNA
Tester (test) cDNA (1st strand)
AAAA
TTTTT
Mix (ratio >35:1) & hybridize
AAAA
TTTT
Hydroxyapatite chromatography
RNA:cDNA hybrids removed
Unhybridized cDNA (differentially expressed) and mRNA
Sepharose CL6B exclusion chromatography
Small cDNA fragments (<450bp)
Enriched, differentially expressed cDNA
or
Produce clones
Label directly and probe library

Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/altered (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/altered population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh

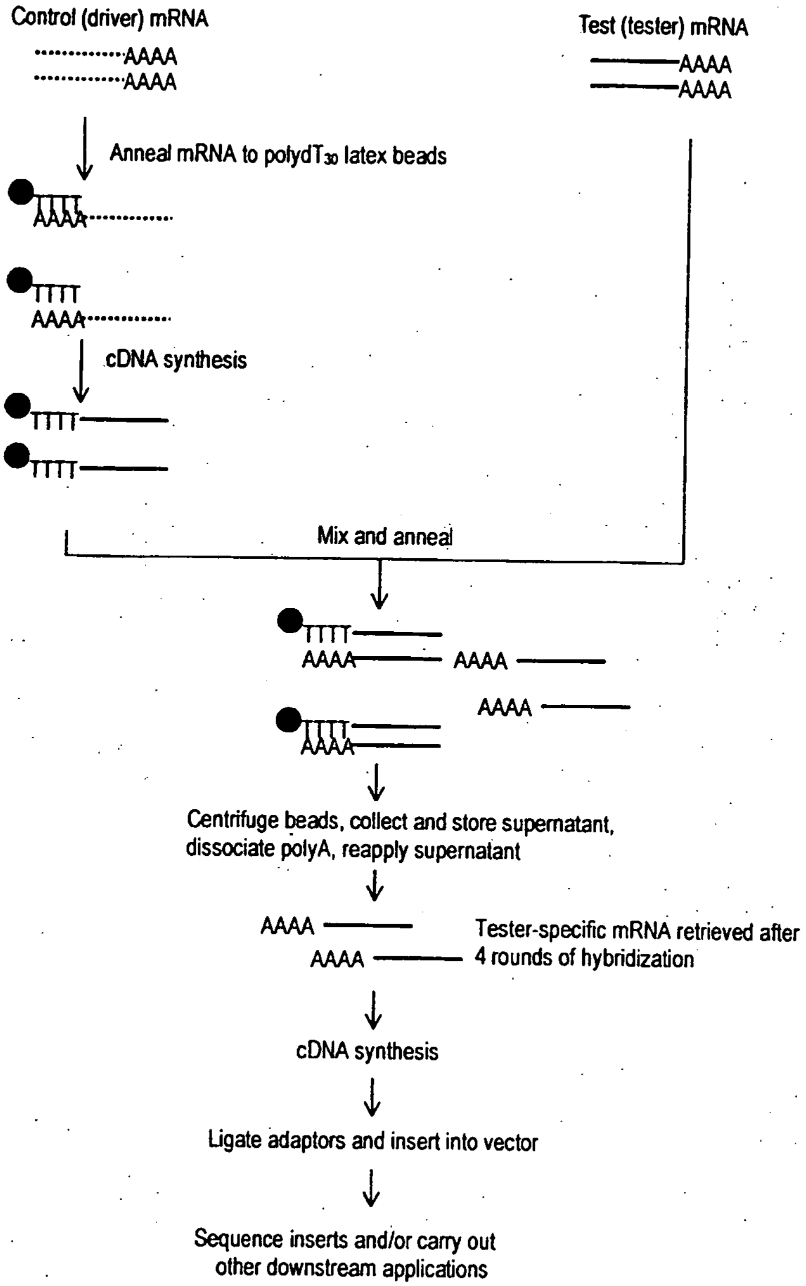

Control (driver) mRNA
AAAA
AAAA
Test (tester) mRNA
AAAA
AAAA
Anneal mRNA to polydT30 latex beads
TTTT
AAAA
TTTT
AAAA
cDNA synthesis
TTTT
TTTT
Mix and anneal
TTTT
AAAA
AAAA
AAAA
TTTT
AAAA
Centrifuge beads, collect and store supernatant, dissociate polyA, reapply supernatant
AAAA
AAAA
Tester-specific mRNA retrieved after 4 rounds of hybridization
cDNA synthesis
Ligate adaptors and insert into vector
Sequence inserts and/or carry out other downstream applications

Figure 2. The use of oligodT$_{30}$ latex to perform subtractive hybridization. mRNA extracted from the control (driver) population is converted to anchored cDNA using polydT oligonucleotides attached to latex beads. mRNA from the treated/altered (tester) population is repeatedly hybridized against an excess of the anchored driver cDNA. The final population of mRNA is tester specific and can be converted into cDNA for cloning and other downstream applications, as described by Hara *et al*. (1991).

control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT_{30}) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT_{30} forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT_{30} population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

Chemical Cross-Linking Subtraction (CCLS)

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 μg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promotor sequence

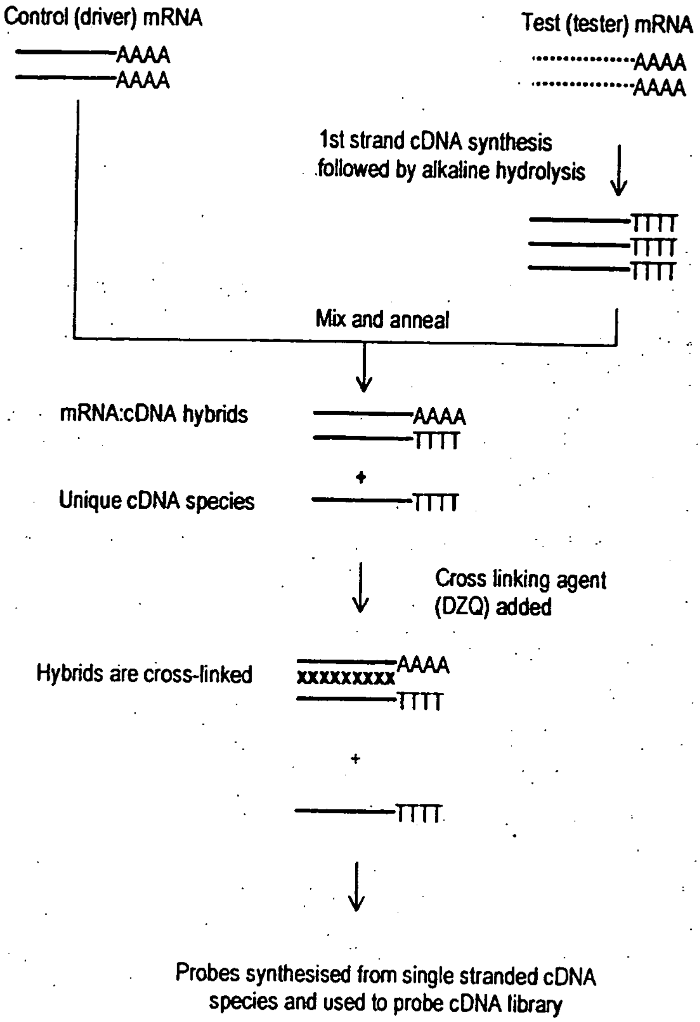


Figure 3. Chemical cross-linking subtraction. Excess driver mRNA is mixed with 1st strand tester cDNA. The common sequences form mRNA:cDNA hybrids which are cross linked with 2,5 diaziridinyl-1,4-benzoquinone (DZQ) and the remaining cDNA sequences are differentially expressed in the tester population. Probes are made from these sequences using Sequenase 2.0 DNA polymerase, which lacks reverse transcriptase activity and, therefore, does not react with the remaining mRNA molecules from the driver. The labelled probes are then used to screen a cDNA library for clones of differentially expressed sequences. Adapted from Walter *et al.* (1996), with permission.

Table 1. The abundance of mRNA species and classes in a typical mammalian cell.

mRNA class	Copies of each species/cell	No. of mRNA species in class	Mean % of each species in class	Mean mass (ng) of each species/μg total RNA
Abundant	12000	4	3.3	1.65
Intermediate	300	500	0.08	0.04
Rare	15	11000	0.004	0.002

Modified from Bertioli *et al.* (1995).

at the 5′ end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

Representational Difference Analysis (RDA)

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5′adaptors at each end of the DNA duplex and can, thus, be filled in at both 3′ ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Shatz used a tester:driver ratio of 1:400, 1:80.000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplifiation and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

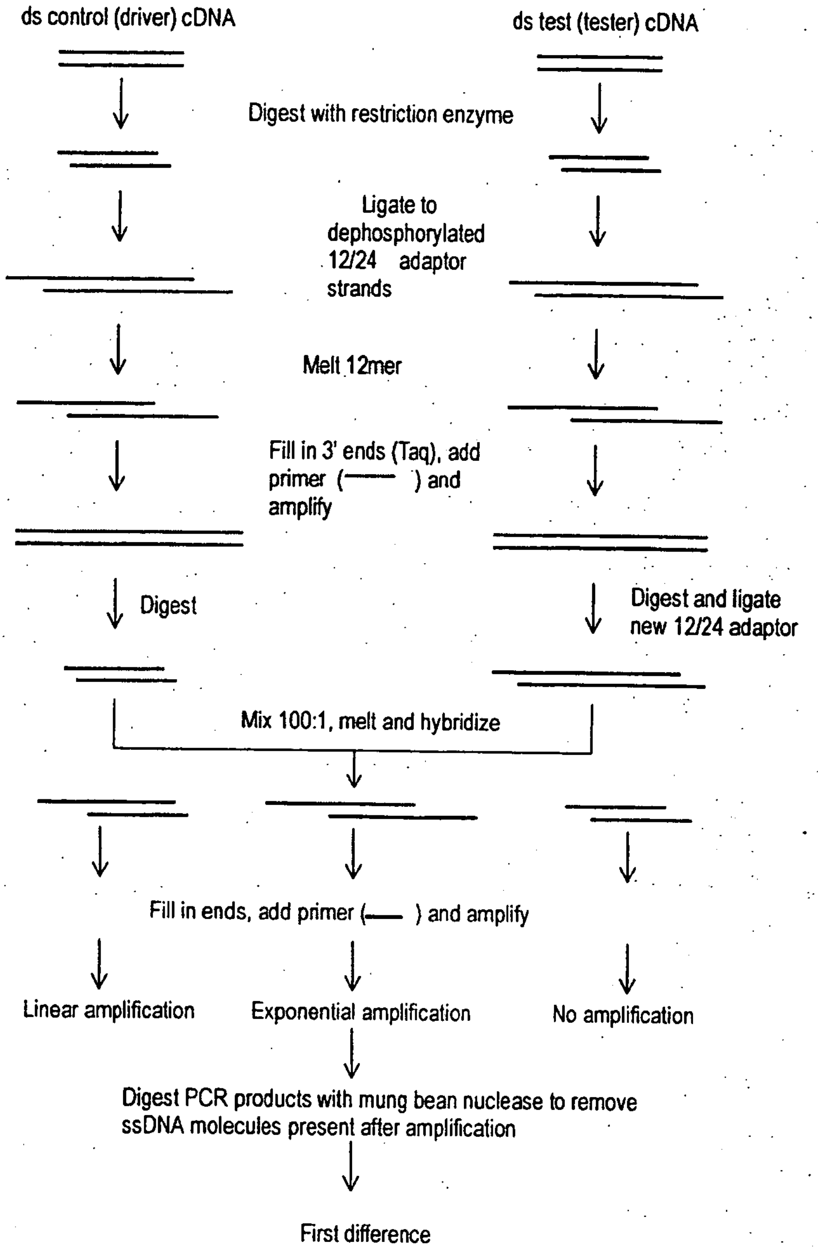

ds control (driver) cDNA
ds test (tester) cDNA
Digest with restriction enzyme
Ligate to dephosphorylated 12/24 adaptor strands
Melt 12mer
Fill in 3' ends (Taq), add primer (——) and amplify
Digest
Digest and ligate new 12/24 adaptor
Mix 100:1, melt and hybridize
Fill in ends, add primer (——) and amplify
Linear amplification
Exponential amplification
No amplification
Digest PCR products with mung bean nuclease to remove ssDNA molecules present after amplification
First difference

Figure 4. The representational difference analysis (RDA) technique. Driver and tester cDNA are digested with a 4-cutter restriction enzyme such as *Dpn*II. The 1st set of 12/24 adaptor strands (oligonucleotides) are ligated to each other and the digested cDNA products. The 12mer is subsequently melted away and the 3'ends filled in using Taq DNA polymerase. Each cDNA population is then amplified using PCR, following which the 1st set of adaptors is removed with *Dpn*II. A second set of 12/24 adaptor strands is then added to the amplified tester cDNA population, after which the tester is hybridized against a large excess of driver. The 12mer adaptors are melted and the 3′ends filled in as before. PCR is carried out with primers identical to the new 24mer adaptor. Thus, the only hybridization products which are exponentially amplified are those which are tester:tester combinations. Following PCR, ssDNA products are removed with mung bean nuclease, leaving the 'first difference product'. This is digested and a third set of 12/24 adaptors added before repeating the subtraction process from the hybridization stage. The process is repeated to the 3rd or 4th difference product, as described by Lisitsyn *et al.* (1993) and Hubank and Schatz (1994).

Suppression PCR Subtractive Hybridization (SSH)

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

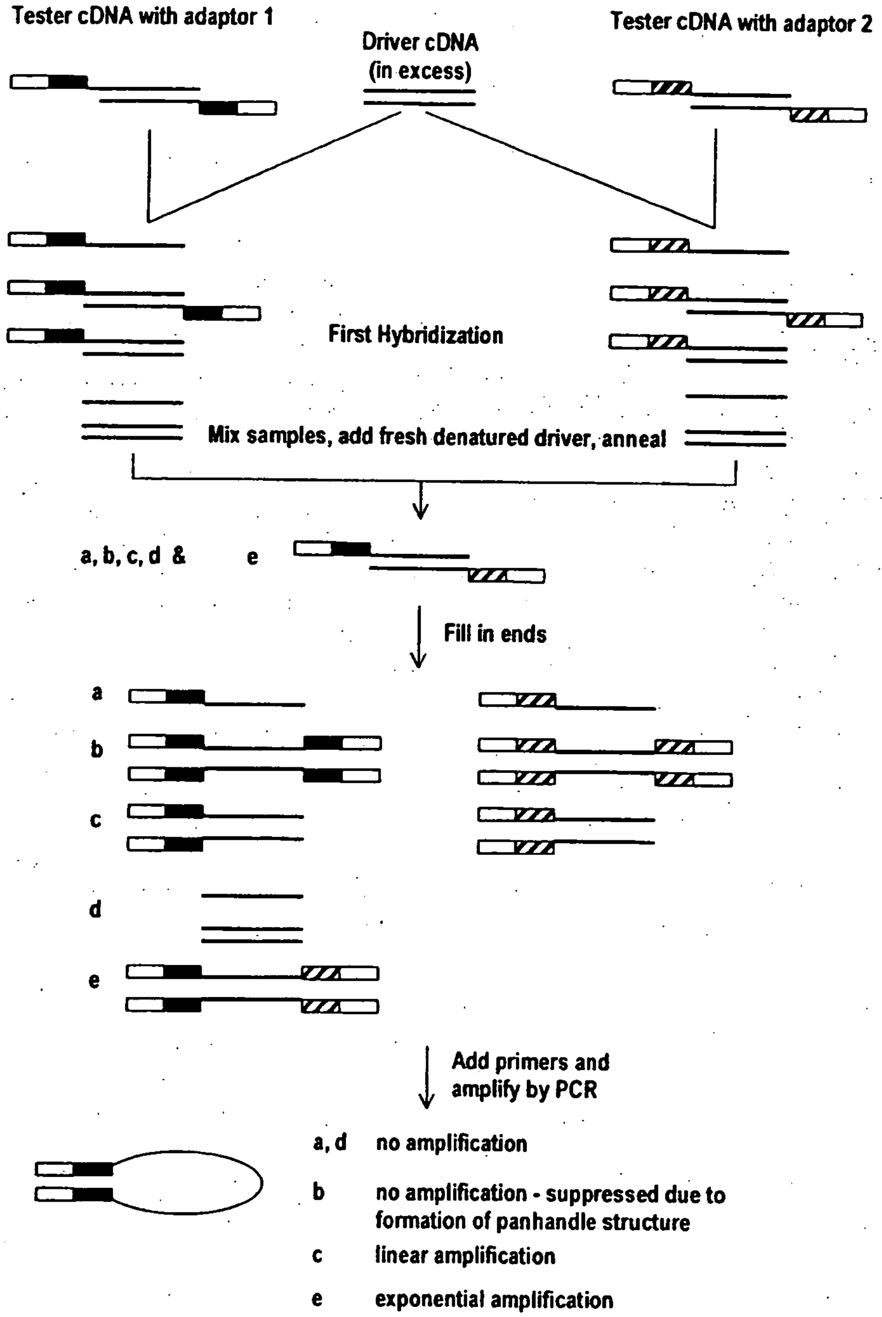

Tester cDNA with adaptor 1
Driver cDNA (in excess)
Tester cDNA with adaptor 2
First Hybridization
Mix samples, add fresh denatured driver, anneal
a, b, c, d &
e
Fill in ends
a
b
c
d
e
Add primers and amplify by PCR
a, d no amplification
b no amplification - suppressed due to formation of panhandle structure
c linear amplification
e exponential amplification

Figure 5. PCR-select cDNA subtraction. In the primary hybridization, an excess of driver cDNA is added to each tester cDNA population. The samples are heat denatured and allowed to hybridize for between 3 and 8 h. This serves two purposes: (1) to equalize rare and abundant molecules; and (2) to enrich for differentially expressed sequences—cDNAs that are not differentially expressed form type c molecules with the driver. In the secondary hybridization, the two primary hybridizations are mixed together without denaturing. Fresh denatured driver can also be added at this point to allow further enrichment of differentially expressed sequences. Type e molecules are formed in this secondary hybridization which are subsequently amplified using two rounds of PCR. The final products can be visualized on an agarose gel, labelled directly or cloned into a vector for downstream manipulation. As described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996), with permission.

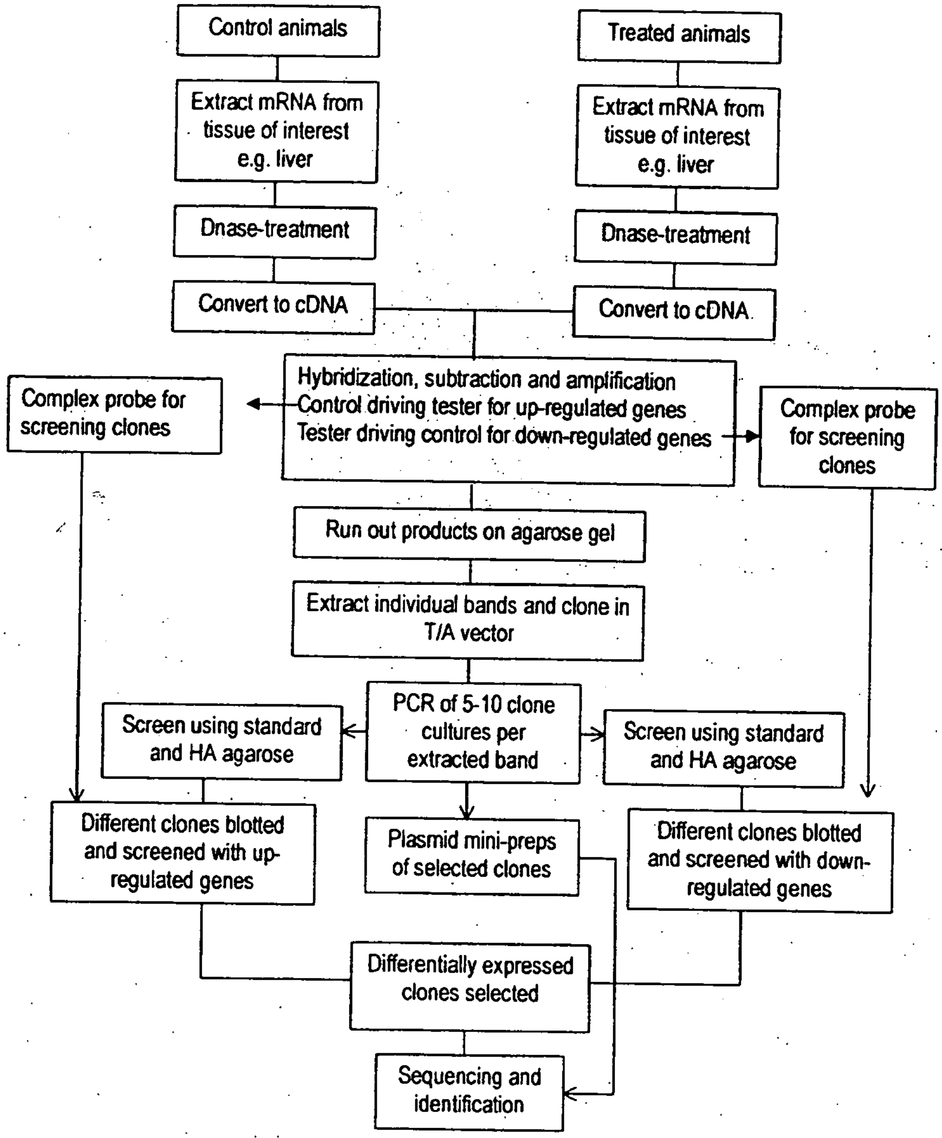

Control animals
Treated animals
Extract mRNA from tissue of interest e.g. liver
Extract mRNA from tissue of interest e.g. liver
Dnase-treatment
Dnase-treatment
Convert to cDNA
Convert to cDNA
Hybridization, subtraction and amplification
Control driving tester for up-regulated genes
Tester driving control for down-regulated genes
Complex probe for screening clones
Complex probe for screening clones
Run out products on agarose gel
Extract individual bands and clone in T/A vector
PCR of 5-10 clone cultures per extracted band
Screen using standard and HA agarose
Screen using standard and HA agarose
Different clones blotted and screened with up-regulated genes
Plasmid mini-preps of selected clones
Different clones blotted and screened with down-regulated genes
Differentially expressed clones selected
Sequencing and identification

Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

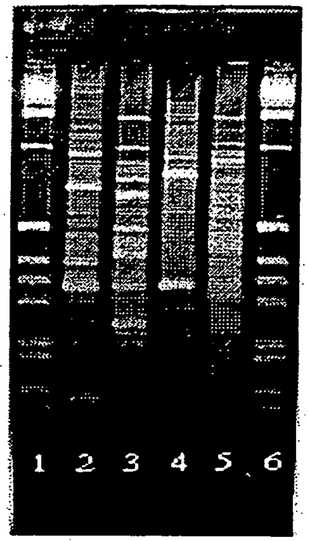


Figure 7. SSH display patterns obtained from rat liver following 3-day treatment with WY-14,643 or phenobarbital. mRNA extracted from control and treated livers was used to generate the differential displays using the PCR-Select cDNA subtraction kit (Clontech). Lane: 1—1kb ladder; 2—genes upregulated following Wy,14–643 treatment; 3—genes downregulated following Wy,14–643 treatment; 4—genes upregulated following phenobarbital treatment; 5—genes downregulated following phenobarbital treatment; 6—1kb ladder. Reproduced from Rockett *et al.* (1997), with permission.

exposure, and an almost complete complement of genes are obtained. For example, the peroxisome proliferator and non-genotoxic hepatocarcinogen Wy,14,643, up-regulates at least 28 genes and down-regulates at least 15 in the rat (a sensitive species) and produces 48 up- and 37 down-regulated genes in the guinea pig, a resistant species (Rockett, Swales, Esda and Gibson, unpublished observations). One of these genes, CD81, was up-regulated in the rat and down-regulated in the guinea pig following Wy-14,643 treatment. CD81 (alternatively named TAPA-1) is a widely expressed cell surface protein which is involved in a large number of cellular processes including adhesion, activation, proliferation and differentiation (Levy *et al.* 1998). Since all of these functions are altered to some extent in the phenomena of hepatomegaly and non-genotoxic hepatocarcinogenesis, it is intriguing, and probably mechanistically-relevant, that CD81 expression is differentially regulated in a resistant and susceptible species. However, the down-side of this approach is that the majority of genes can be sequenced and matched to database sequences, but the latter are predominantly expressed sequence tags or genes of completely unknown function, thus partially obscuring a realistic overall assessment of the critical genes of genuine biological interest. Notwithstanding the lack of complete funtional identification of altered gene expression, such gene profiling studies essentially provides a 'molecular fingerprint' in response to xenobiotic challenge, thereby serving as a mechanistically-relevant platform for further detailed investigations.

Differential Display (DD)

Originally described as 'RNA fingerprinting by arbitrarily primed PCR' (Liang and Pardee 1992) this method is now more commonly referred to as 'differential

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 *H. sapiens* (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA Sulfated glycoprotein
15 (600)	92.9%	Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobulin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNa

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot anaylsis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopoxin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	*M.musculus* Cl inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	*M. musculus* Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT *M. musculus* (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML *M. musculus* (EST)
	Clone 2 82:0%	Soares p3NMF 19.5 *M. musculus* (EST)
10 (550)	73.8%	Soares mouse NML *M. musculus* (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 *H. sapiens* (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 *M. musculus* (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	*R. norvegicus* RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simiply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one display compared to the other, are differentially expressed and may be recovered for further characterization. One advantage of this system is the speed with which it can be carried out—2 days to obtain a display and as little as a week to make and identify clones.

Two commonly used variations are based on different methods of priming the reverse transcription step (figure 8). One is to use an oligo dT with a 2-base 'anchor' at the 3'-end, e.g. 5′ (dT_{11})CA 3′ (Liang and Pardee 1992). Alternatively, an arbitrary primer may be used for 1st strand cDNA synthesis (Welsh *et al.* 1992). This variant of RNA fingerprinting has also been called 'RAP' (RNA Arbitrarily Primed)-PCR. One advantage of this second approach is that PCR products may be derived from anywhere in the RNA, including open reading frames. In addition, it can be used for mRNAs that are not polyadenylated, such as many bacterial mRNAs (Wong and McClelland 1994). In both cases, following reverse transcription and denaturation, second strand cDNA synthesis is carried out with an arbitrary primer (*arbitrary* primers have a single base at each position, as compared to *random* primers, which contain a mixture of all four bases at each position). The resulting PCR, thus, produces a series of products which, depending on the system (primer length and composition, polymerase and gel system), usually includes 50–100 products per primer set (Band and Sager 1989). When a combination of different dT-anchors and arbitrary primers are used, almost all mRNA species from a cell can be amplified. When the cDNA products from two different populations are analysed side by side on a polyacrylamide gel, differences in expression can be identified and the appropriate bands recovered for cloning and further analysis.

Although DD is perhaps the most popular approach used today for identifying differentially expressed genes, it does suffer from several perceived disadvantages:

(1) It may have a strong bias towards high copy number mRNAs (Bertioli *et al.* 1995), although this has been disputed (Wan *et al.* 1996) and the isolation of very low abundance genes may be achieved in certain circumstances (Guimeraes *et al.* 1995a).

(2) The cDNAs obtained often only represent the extreme 3′ end of the mRNA (often the 3′-untranslated region), although this may not always be the case (Guimeraes *et al.* 1995a). Since the 3′ end is often not included in Genbank and shows variation between organisms, cDNAs identified by DD cannot always be matched with their genes, even if they have been identified.

(3) The pattern of differential expression seen on the display often cannot be reproduced on Northern blots, with false positives arising in up to 70% of cases (Sun *et al.* 1994). Some adaptations have been shown to reduce false positives, including the use of two reverse transcriptases (Sung and Denman 1997), comparison of uninduced and induced cells over a time course (Burn *et al.* 1994) and comparison of DDPCR-products from two uninduced and two induced lines (Sompayrac *et al.* 1995). The latter authors also reported that the use of cytoplasmic RNA rather then total RNA reduces false positives arising from nuclear RNA that is not transported to the cytoplasm.

Further details of the background, strengths and weaknesses of the DD technique can be obtained from a review by McClelland *et al.* (1996) and from articles by Liang *et al.* (1995) and Wan *et al.* (1996).

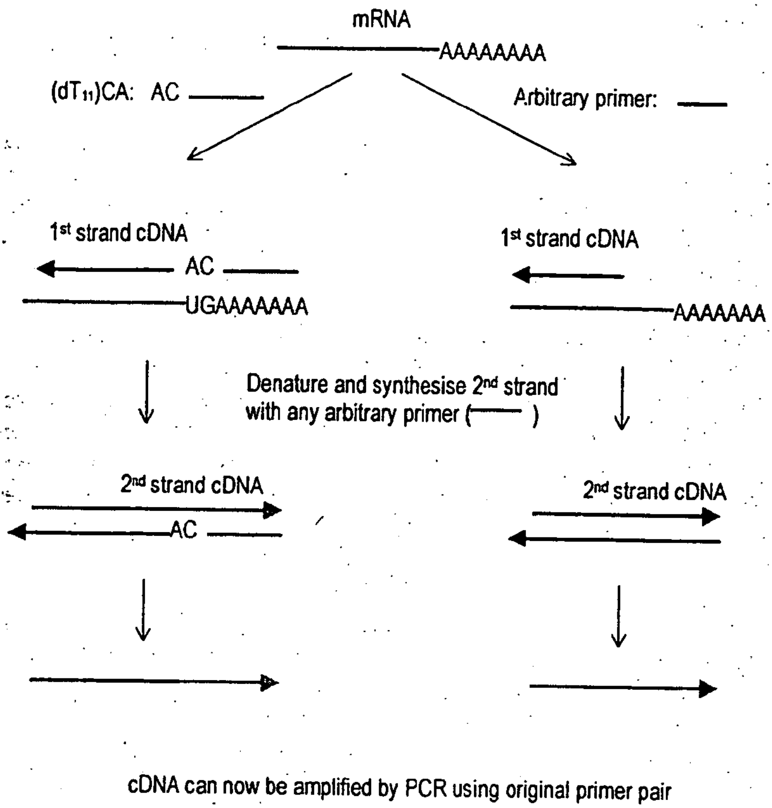

mRNA
AAAAAAAA
(dT11)CA: AC
Arbitrary primer:
1st strand cDNA
AC
UGAAAAAAA
1st strand cDNA
AAAAAAA
Denature and synthesise 2nd strand with any arbitrary primer ( )
2nd strand cDNA
AC
2nd strand cDNA
cDNA can now be amplified by PCR using original primer pair

Figure 8. Two approaches to differential display (DD) analysis. 1st strand synthesis can be carried out either with a polydT$_{11}$ NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polydT primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1st strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2nd strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2nd strand may also hybridize to the 1st strand cDNA in a number of different places, several different 2nd strand products may be obtained from one binding point of the 1st strand primer. Following 2nd strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

Restriction endonuclease-facilitated analysis of gene expression

Serial Analysis of Gene Expression (SAGE)

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatonation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3′ ends of the cDNA population are captured with streptavidin beads. The captured population is split into two and different adaptors ligated to the 5′ends of each group. Incorporated into the adaptors is a recognition sequence for a type IIS restriction enzyme—one which cuts DNA at a defined distance (< 20 bp) from its recognition sequence. Hence, following digestion of each captured cDNA population with the IIS enzyme, the adaptors plus a short piece of the captured cDNA are released. The two populations are then ligated and the products amplified. The amplified products are cleaved with the original anchoring enzyme, religated (concatomers are formed in the process) and cloned. The advantage of this system is that hundreds of gene tags can be identified by sequencing only a few clones. Furthermore, the number of times a given transcript is identified is a quantitative measurement of that gene's abundance in the original population, a feature which facilitates identification of differentially expressed genes in different cell populations.

Some disadvantages of SAGE analysis include the technical difficulty of the method, a large amount of accurate sequencing is required, biased towards abundant mRNAs, has not been validated in the pharmaco/toxicogenomic setting and has only been used to examine well known tissue differences to date.

Gene Expression Fingerprinting (GEF)

A different capture/restriction digest approach for isolating differentially expressed genes has been described by Ivanova and Belyavsky (1995). In this method, RNA is converted to cDNA using biotinylated oligo(dT) primers. The cDNA population is then digested with a specific endonuclease and captured with magnetic streptavidin microbeads to facilitate removal of the unwanted 5′digestion products. The use of restricted 3′-ends alone serves to reduce the complexity of the cDNA fragment pool and helps to ensure that each RNA species is represented by not more than one restriction product. An adaptor is ligated to facilitate subsequent amplification of the captured population. PCR is carried out with one adaptor-specific and one biotinylated polydT primer. The reamplified population is recaptured and the non-biotinylated strands removed by alkaline dissociation. The non-biotinylated strand is then resynthesized using a different adaptor-specific primer in the presence of a radiolabelled dNTP. The labelled immobilized 3′cDNA ends are next sequentially treated with a series of different restriction endonucleases and the products from each digestion analysed by PAGE. The result is a fingerprint composed of a number of ladders (equal to the number of sequential digests used). By comparing test versus control fingerprints, it is possible to identify differentially expressed products which can then be isolated from the gel and cloned. The advantages of this procedure are that it is very robust and reproducible, and the authors estimate that 80–93% of cDNA molecules are involved in the final fingerprint. The disadvantage is that polyacrylamide gels can rarely resolve more than 300–400 bands, which compares poorly to the 1000 or more which are estimated to be produced in an average experiment. The use of 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991) may help to overcome this problem.

A similar method for displaying restriction endonuclease fragments was later described by Prashar and Weissman (1996). However, instead of sequential digestion of the immobolized 3'-terminal cDNA fragments, these authors simply compared the profiles of the control and treated populations without further manipulation.

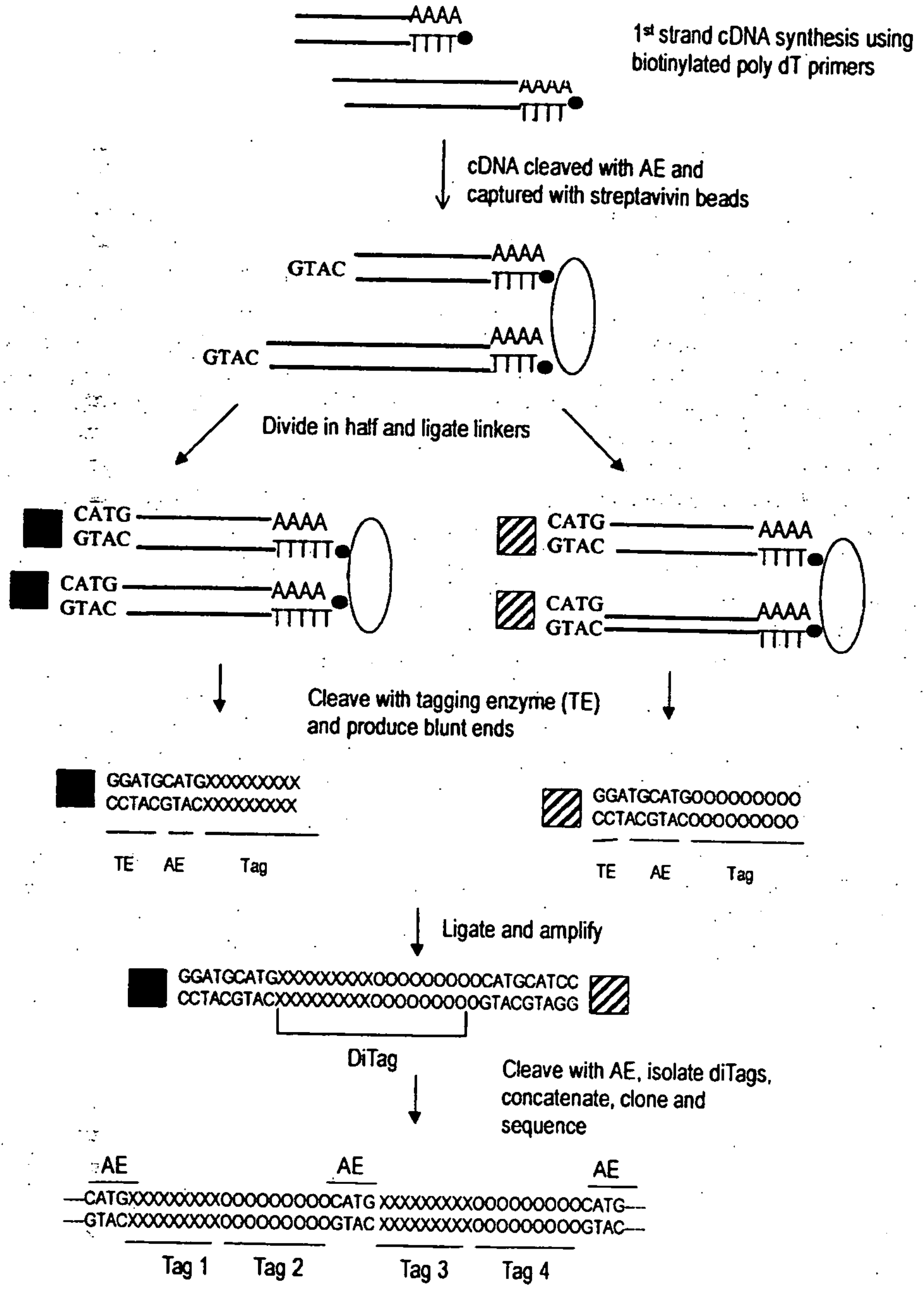

AAAA
TTTT
1st strand cDNA synthesis using biotinylated poly dT primers
AAAA
TTTT
cDNA cleaved with AE and captured with streptavivin beads
GTAC
AAAA
TTTT
GTAC
AAAA
TTTT
Divide in half and ligate linkers
CATG
GTAC
AAAA
TTTTT
CATG
GTAC
AAAA
TTTTT
CATG
GTAC
AAAA
TTTT
CATG
GTAC
AAAA
TTTT
Cleave with tagging enzyme (TE) and produce blunt ends
GGATGCATGXXXXXXXXX
CCTACGTACXXXXXXXXX
TE
AE
Tag
GGATGCATGOOOOOOOOO
CCTACGTACOOOOOOOOO
TE
AE
Tag
Ligate and amplify
GGATGCATGXXXXXXXXXOOOOOOOOOCATGCATCC
CCTACGTACXXXXXXXXXOOOOOOOOOGTACGTAGG
DiTag
Cleave with AE, isolate diTags, concatenate, clone and sequence
AE
AE
AE
—CATGXXXXXXXXXOOOOOOOOOCATG XXXXXXXXXOOOOOOOOOCATG—
—GTACXXXXXXXXXOOOOOOOOOGTAC XXXXXXXXXOOOOOOOOOGTAC—
Tag 1
Tag 2
Tag 3
Tag 4

Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3'ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the ditags isolated from the linkers using PAGE. The ditags are then ligated (during which process, concatenization occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

DNA arrays

'Open' differential display systems are cumbersome in that it takes a great deal of time to extract and identify candidate genes and then confirm that they are indeed up- or down-regulated in the treated compared to the control tissue. Normally, the latter process is carried out using Northern blotting or RT-PCR. Even so, each of the aforementioned steps produce a bottleneck to the ultimate goal of rapid analysis of gene expression. These problems will likely be addressed by the development of so-called DNA arrays (e.g. Gress *et al.* 1992, Zhao *et al.* 1995, Schena *et al.* 1996), the introduction of which has signalled the next era in differential gene expression analysis. DNA arrays consist of a gridded membrane or glass 'chips' containing hundreds or thousands of DNA spots, each consisting of multiple copies of part of a known gene. The genes are often selected based on previously proven involvement in oncogenesis, cell cycling, DNA repair, development and other cellular processes. They are usually chosen to be as specific as possible for each gene and animal species. Human and mouse arrays are already commercially available and a few companies will construct a personalized array to order, for example Clontech Laboratories and Research Genetics Inc. The technique is rapid in that hundreds or even thousands of genes can be spotted on a single array, and that mRNA/cDNA from the test populations can be labelled and used directly as probe. When analysed with appropriate hardware and software, arrays offer a rapid and quantitative means to assess differences in gene expression between two cell populations. Of course, there can only be identification and quantitation of those genes which are in the array (hence the term 'closed' system). Therefore, one approach to elucidating the molecular mechanisms involved in a particular disease/development system may be to combine an open and closed system—a DNA array to directly identify and quantitate the expression of known genes in mRNA populations, and an open system such as SSH to isolate unknown genes which are differentially expressed.

One of the main advantages of DNA arrays is the huge number of gene fragments which can be put on a membrane—some companies have reported gridding up to 60000 spots on a single glass 'chip' (microscope slide). These high density chip-based micro-arrays will probably become available as mass-produced off-the-shelf items in the near future. This should facilitate the more rapid determination of differential expression in time and dose-response experiments. Aside from their high cost and the technical complexities involved in producing and probing DNA arrays, the main problem which remains, especially with the newer micro-array (gene-chip) technologies, is that results are often not wholly reproducible between arrays. However, this problem is being addressed and should be resolved within the next few years.

EST databases as a means to identify differentially expressed genes

Expressed sequence tags (ESTs) are partial sequences of clones obtained from cDNA libraries. Even though most ESTs have no formal identity (putative identification is the best to be hoped for), they have proven to be a rapid and efficient means of discovering new genes and can be used to generate profiles of gene-expression in specific cells. Since they were first described by Adams *et al.* (1991), there has been a huge explosion in EST production and it is estimated that there are now well over a million such sequences in the public domain, representing over half of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatzis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at http://www.tigr.org) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BAC)s, or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatzis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

Problems and potential of differential expression techniques

The holistic or single cell approach?

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new microdisection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: http://www.ncbi.nlm.nih.gov/ncicgap/intro.html). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells, e.g. fluorescence activated cell sorting (FACS) (Dunbar *et al.* 1998, Kas-Deelen *et al.* 1998) and magnetic bead technology (Richard *et al.* 1998, Rogler *et al.* 1998).

However, those taking a holistic approach may consider this issue unimportant. There is an equally appropriate view that all those genes showing altered expression within a compromized tissue should be taken into consideration. After all, since all tissues are complex mixes of different, interacting cell types which intimately regulate each other's growth and development, it is clear that each cell type could in some way contribute (positively or negatively) towards the molecular mechanisms which lie behind responses to external stimuli or neoplastic growth. It is perhaps then more informative to carry out differential display experiments using *in vivo* as opposed to *in vitro* models, where uniform populations of identical cells probably represent a partial, skewed or even inaccurate picture of the molecular changes that occur.

The incidence and possible implications of inter-individual biological variation should be considered in any approach where whole animal models are being used. It is clear that individuals (humans and animals) respond in different ways to identical stimuli. One of the best characterized examples is the debrisoquine oxidation polymorphism, which is mediated by cytochrome CYP2D6 and determines the pharmacokinetics of many commonly prescribed drugs (Lennard 1993, Meyer and Zanger 1997). The reasons for such differences are varied and complex, but allelic variations, regulatory region polymorphisms and even physical and mental health can all contribute to observed differences in individual responses. Careful thought should, therefore, be given to the specific objectives of the study and to the possible value of pooling starting material (tissue/mRNA). The effect of this can be beneficial through the ironing out of exaggerated responses and unimportant minor fluctuations of (mechanistically) irrelevant genes in individual animals, thus providing a clearer overall picture of the general molecular mechanisms of the response. However, at the same time such minor variations may be of utmost importance in deciding the ability of individual animals to succumb to or resist the effects of a given chemical/disease.

How efficient are differential expression techniques at recovering a high percentage of differentially expressed genes?

A number of groups have produced experimental data suggesting that mammalian cells produce between 8000–15000 different mRNA species at any one time (Mechler and Rabbitts 1981, Hedrick *et al.* 1984, Bravo 1990), although figures as high as 20–30000 have also been quoted (Axel *et al.* 1976). Hedrick *et al.* (1984) provided evidence suggesting that the majority of these belong to the rare abundance class. A breakdown of this abundance distribution is shown in table 1.

When the results of differential display experiments have been compared with data obtained previously using other methods, it is apparent that not all differentially expressed mRNAs are represented in the final display. In particular, rare messages (which, importantly, often include regulatory proteins) are not easily recovered using differential display systems. This is a major shortcoming, as the majority of mRNA species exist at levels of less than 0.005% of the total population (table 1). Bertioli *et al.* (1995) examined the efficiency of DD templates (heterogeneous mRNA populations) for recovering rare messages and were unable to detect mRNA species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000× smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon-γ-stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3′ restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true positives), it is still not clear if such adaptations are practically effective—proving efficiency by spiking with a known amount of limited numbers of artificial construct(s) is one thing, but isolating a high percentage of the rare messages already present in an mRNA population is another. Of course, some models will genuinely produce only a small number of differentially expressed genes. In addition, there are also technical problems that can reduce efficiency. For example, mRNAs may have an unusual primary structure that effectively prevents their amplification by PCR-based systems. In addition, it is known that under certain circumstances not all mRNAs have 3′polyA sites. For example, during *Xenopus* development, deadenylation is used as a means to stabilize RNAs (Voeltz and Steitz 1998), whilst preferential deadenylation may play a role in regulating Hsp70 (and perhaps, therefore, other stress protein) expression in *Drosophila* (Dellavalle *et al.* 1994). The presence of deadenylated mRNAs would clearly reduce the efficiency of systems utilizing a polydT reverse transcription step. The efficiency of any system also depends on the quality of the starting material. All differential display techniques use mRNA as their target material. However, it is difficult to isolate mRNA that is completely free of ribosomal RNA. Even if polydT primers are used to prime first strand cDNA synthesis, ribosomal RNA is often transcribed to some degree (Clontech PCR-Select cDNA Subtraction kit user manual). It has been shown, at least in the case of SSH, that a high rRNA:mRNA ratio can lead to inefficient subtractive hybridization (Clontech PCR-Select cDNA Subtraction kit user manual), and there is no reason to suppose that it will not do likewise in other SH approaches. Finally, those techniques that utilise a presubtraction amplification step (e.g. RDA) may present a skewed representation since some sequences amplify better than others.

Of course, probably the most important consideration is the temporal factor. It is clear that any given differential display experiment can only interrogate a cell at one point in time. It may well be that a high percentage of the genes showing altered expression at that time are obtained. However, given that disease processes and responses to environmental stimuli involve dynamic cascades of signalling, regulation, production and action, it is clear that all those genes which are switched on/off at different times will not be recovered and, therefore, vital information may well be missed. It is, therefore, imperative to obtain as much information about the model system beforehand as possible, from which a strategy can be derived for targeting specific time points or events that are of particular interest to the investigator. One way of getting round this problem of single time point analysis is to conduct the experiment over a suitable time course which, of course, adds substantially to the amount of work involved.

How sensitive are differential expression technologies?

There has been little published data that addresses the issue of how large the change in expression must be for it to permit isolation of the gene in question with the various differential expression technologies. Although the isolation of genes whose expression is changed as little as 1.5-fold has been reported using SSH (Groenink and Leegwater 1996), it appears that those demonstrating a change in excess of 5-fold are more likely to be picked up. Thus, there is a 'grey zone' in between where small changes could fade in and out of isolation between experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

Resolution and visualization of differential expression products

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hessle, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

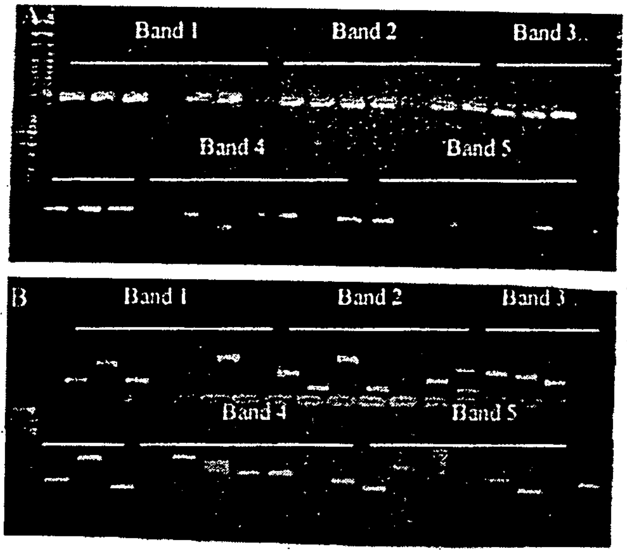


Figure 10. Discrimination of clones of identical/nearly identical size using HA-red. Bands of decreasing size (1–5) were extracted from the final display of a suppression subtractive hybridization experiment and cloned. Seven colonies were picked at random from each cloned band and their inserts amplified using PCR. The products were run on two gels, (A) a high resolution 2% agarose gel, and (B) a high resolution 2% agarose gel containing 1 U/ml HA-red. With few exceptions, all the clones from each band appear to be the same size (gel A). However, the presence of HA-red (gel B), which separates identically-sized DNA fragments based on the percentage of GC within the sequence, clearly indicates the presence of different gene species within each band. For example, even though all five re-amplified clones of band 1 appear to be the same size, at least four different gene species are represented.

in a similar gel containing one of the HA-stains. The standard gel should indicate any gross size differences, whilst the HA-stained gel should separate otherwise unresolvable species (on standard AGE) according to their base content. Geisinger *et al.* (1997) reported successful use of this approach for identifying DD-derived clones. Figure 10 shows such an experiment carried out in this laboratory on clones obtained from a band extracted from an SSH display.

An alternative approach is to carry out a 2-D analysis of the differential display products. In this approach, size-based separation is first carried out in a standard agarose gel. The gel slice containing the display is then extracted and incorporated in to a HA gel for resolution based on AT/GC content.

Of course, one should always consider the possibility of there being different gene species which are the same size and have the same GC/AT content. However, even these species are not unresolvable given some effort—again, one might use SSCP, or perhaps a denaturing gradient gel electrophoresis (DGGE) or temperature gradient field electrophoresis (TGGE) approach to resolve the contents of a band, either directly on the extracted band (Suzuki *et al.* 1991) or on the reamplified product.

The requirement of some differential display techniques to visualize large numbers of products (e.g. DD and GEF) can also present a problem in that, in terms of numbers, the resolution of PAGE rarely exceeds 300–400 bands. One approach to overcoming this might be to use 2-D gels such as those described by Uitterlinden *et al.* (1989) and Hatada *et al.* (1991).

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to overstain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

The possible use of 'microfingerprinting' to reduce complexity

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 + bp on the same gel, which leads to compromize in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to included certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

An alternative approach to microfingerprinting is to examine altered expression in specific families of genes through careful selection of PCR primers and/or post-reaction analysis. Stress genes, growth factors and/or their receptors, cell cycling genes, cytochromes P450 and regulatory proteins might be considered as candidates for analysis in this way. Indeed, some off-the-shelf DNA arrays (e.g. Clontech's Atlas cDNA Expression Array series) already anticipated this to some degree by grouping together genes involved in different responses e.g. apoptosis, stress, DNA-damage response etc.

Screening

False positives

The generation of false positives has been discussed at length amongst the differential display community (Liang *et al.* 1993, 1995, Nishio *et al.* 1994, Sun *et al.* 1994, Sompayrac *et al.* 1995). The reason for false positives varies with the technique being used. For instance, in RDA, the use of adaptors which have not been HPLC purified can lead to the production of false positives through illegitimate ligation events (O'Neill and Sinclair 1997), whilst in DD they can arise through PCR artifacts and illegitemate transcription of rRNA. In SH, false positives appear to be derived largely from abundant gene species, although some may arise from cDNA/mRNA species which do not undergo hybridization for technical reasons.

A quick screening of putative differentially expressed clones can be carried out using a simple dot blot approach, in which labelled first strand probes synthesized from tester and driver mRNA are hybridized to an array of said clones (Hedrick *et al.* 1984, Sakaguchi *et al.* 1986). Differentially expressed clones will hybridize to tester probe, but not driver. The disadvantage of this approach is that rare species may not generate detectable hybridization signals. One option for those using SSH is to screen the clones using a labelled probe generated from the subtracted cDNA from which it was derived, and with a probe made from the reverse subtraction reaction (ClonTechniques 1997a). Since the SSH method enriches rare sequences, it should be possible to confirm the presence of clones representing low abundance genes. Despite this quick screening step, there is still the need to go back to the original mRNA and confirm the altered expression using a more quantitative approach. Although this may be achieved using Northern blots, the sensitivity is poor by today's high standards and one must rely on PCR methods for accurate and sensitive determinations (see below).

Sequence analysis

The majority of differential display procedures produce final products which are between 100 and 1000bp in size. However, this may considerably reduce the size of the sequence for analysis of the DNA databases. This in turn leads to a reduced confidence in the result—several families of genes have members whose DNA sequences are almost identical except in a few key stretches, e.g. the cytochrome P450 gene superfamily (Nelson *et al.* 1996). Thus, does the clone identified as being almost identical to gene X_0 really come from that gene, or its brother gene X_1 or its as yet undiscovered sister X_2? For example, using SSH, part of a gene was isolated, which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

(1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
(2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identitiy between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using mulitchannel pipettes, 96 +-well plates and appropriate thermal cycling technology. Whilst quantitative analysis is more desirable, being more accurate and without reliance on an internal standard, the money and time needed to develop a competitor molecule is often excessive, especially when one might be examining tens or even hundreds of gene species. The use of semi-quantitative analysis is simpler, although still relatively involved. One must first of all choose an internal standard that does not change in the test cells compared to the controls. Numerous reference genes have been tried in the past, for example interferon-gamma (IFN-γ, Frye *et al.* 1989), β-actin (Heuval *et al.* 1994), glyceraldehyde-3-phosphate dehydrogenase (GAPDH, Wong *et al.* 1994), dihydrofolate reductase (DHFR, Mohler and Butler 1991), β-2-microglobulin (β-2-m, Murphy *et al.* 1990), hypoxanthine phosphoribosyl transferase (HPRT, Foss *et al.* 1998) and a number of others (ClonTechniques 1997b). Ideally, an internal standard should not change its level of expression in the cell regardless of cell age, stage in the cell cycle or through the effects of external stimuli. However, it has been shown on numerous occasions that the levels of most housekeeping genes currently used by the research community do in fact change under certain conditions and in different tissues (ClonTechniques 1997b). It is imperative, therefore, that preliminary experiments be carried out on a panel of housekeeping genes to establish their suitability for use in the model system.

Interpretation of quantitative data must also be treated with caution. By comparing the lists of genes identified by differential expression one can perhaps gain insight into why two different species react in different ways to external stimuli. For example, rats and mice appear sensitive to the non-genotoxic effects of a wide range of peroxisome proliferators whilst Syrian hamsters and guinea pigs are largely resistant (Orton *et al.* 1984, Rodricks and Turnbull 1987, Lake *et al.* 1989, 1993, Makowska *et al,* 1992). A simplified approach to resolving the reason(s) why is to compare lists of up- and down-regulated genes in order to identify those which are expressed in only one species and, through background knowledge of the effects of the said gene, might suggest a mechanism of facilitated non-genotoxic carcinogenesis or protection. Of course, the situation is likely to be far more complex. Perhaps if there were one key gene protecting guinea pig from non-genotoxic effects and it was upregulated 50 times by PPs, the same gene might only be up-regulated five times in the rat. However, since both were noted to be upregulated, the importance of the gene may be overlooked. Just to complicate matters, a large change in expression does not necessarily mean a biologically important change. For example, what is the true relevance of gene Y which shows a 50-fold increase after a particular treatment, and gene Z which shows only a 5-fold increase? If one examines the literature one may find that historically, gene Y has often been shown to be up-regulated 40–60-fold by a number of unrelated stimuli—in light of this the 50-fold increase would appear less significant. However, the literature may show that gene Z has never been recorded as having more than doubled in expression—which makes your 5-fold increase all the more exciting. Perhaps even more interesting is if that same 5-fold increase has only been seen in related neoplasms or following treatment with related chemicals.

Problems in using th differ ntial display approach

Differential display technology originally held promise of an easily obtainable 'fingerprint' of those genes which are up- or down-regulated in test animals/cells in a developmental process or following exposure to given stimuli. However, it has become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

Conclusions

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the carcinogenic effect. Whilst differential display technology cannot hope to answer these questions, it does provide a springboard from which identification, regulatory and functional studies can be launched. Understanding the molecular mechanism of cellular responses is almost impossible without knowing the regulation and function of those genes and their condition (e.g. mutated). In an abstract sense, differential display can be likened to a still photograph, showing details of a fixed moment in time. Consider the Historian who knows the outcome of a battle and the placement and condition of the troops before the battle commenced, but is asked to try and deduce how the battle progressed and why it ended as it did from a few still photographs—an impossible task. In order to understand the battle, the Historian must find out the capabilities and motivation of the soldiers and their commanding officers, what the orders were and whether they were obeyed. He must examine the terrain, the remains of the battle and consider the effects the prevailing weather conditions exerted. Likewise, if mechanistic answers are to be forthcoming, the scientist must use differential display in combination with other techniques, such as knockout technology, the analysis of cell signalling pathways, mutation analysis and time and dose response analyses. Although this review has emphasized the importance of differential gene profiling, it should not be considered in isolation and the full impact of this approach will be strengthened if used in combination with functional genomics and proteomics (2-dimensional protein gels from isoelectric focusing and subsequent SDS electrophoresis and virtual 2D-maps using capillary electrophoresis). Proteomics is attracting much recent attention as many of the changes resulting in differential gene expression do not involve changes in mRNA levels, as decribed extensively herein, but rather protein–protein, protein–DNA and protein phosphorylation events which would require functional genomics or proteomic technologies for investigation.

Despite the limitations of differential display technology, it is clear that many potential applications and benefits can be obtained from characterizing the genetic changes that occur in a cell during normal and disease development and in response to chemical or biological insult. In light of functional data, such profiling will provide a 'fingerprint' of each stage of development or response, and in the long term should help in the elucidation of specific and sensitive biomarkers for different types of chemical/biological exposure and disease states. The potential medical and therapeutic benefits of understanding such molecular changes are almost immeasurable. Amongst other things, such fingerprints could indicate the family or even specific type of chemical an individual has been exposed to plus the length and/or acuteness of that exposure, thus indicating the most prudent treatment. They may also help uncover differences in histologically identical cancers, provide diagnostic tests for the earliest stages of neoplasia and, again, perhaps indicate the most efficacious treatment.

The Human Genome Project will be completed early in the next century and the DNA sequence of all the human genes will be known. The continuing development and evolution of differential gene expression technology will ensure that this knowledge contributes fully to the understanding of human disease processes.

Acknowledgements

We acknowledge Drs Nick Plant (University of Surrey), Sally Darney and Chris Luft (US EPA at RTP) for their critical analysis of the manuscript prior to submission. This manuscript has been reviewed in accordance with the policy of the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R. and Ventor, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

An, G., Luo, G., Veltri, R. W. and O'Hara, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342–346.

Axel, R., Feigelson, P. and Schultz, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247–254.

Band, V. and Sager, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the Naional Academy of Sciences, USA*, **86**,1249–1253.

Bauer, D., Muller, H., Reich, J., Riedel, H., Ahrenkiel, V., Warthoe, P. and Strauss, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272–4280.

Bertioli, D. J., Schlichter, U. H. A., Adams, M. J., Burrows, P. R., Steinbiss, H.-H. and Antoniw, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520–4523.

Bravo, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37–46.

Burn, T. C., Petrovick, M. S., Hohaus, S., Rollins, B. J. and Tenen, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776–2783.

Cao, J., Cai, X., Zheng, L., Geng, L., Shi, Z., Pao, C. C. and Zheng, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447–451.

Cassidy, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13–20.

Chang, G. W. and Terzaghi-Howe, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445–4452.

Chen, J., Schwartz, D. A., Young, T. A., Norris, J. S. and Yager, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783–2786.

Chen, J. J. W. and Peck, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793–794.

Clon Techniques, 1997a, PCR-Select Differential Screening Kit—the nextstep after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18–19.

Clon Techniques, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15–16.

Davis, M. M., Cohen, D. I., Nielsen, E. A., Steinmetz, M., Paul, W. E. and Hood, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194–2198.

Dellavalle, R. P., Peterson, R. and Lindquist, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in Drosophila melanogaster. *Molecular and Cell Biology*, **14**, 3646–3659.

DeRisi, J. L., Vashwanath, R. L. and Brown, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Diatchenko, L., Lau, Y.-F. C., Campbell, A. P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, K., Gurskaya, N., Sverdlov, E. D. and Siebert, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025–6030.

Dogra, S. C., Whitelaw, M. L. and May, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1–9.

Duguid, J. R. and Dinauer, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789–2792.

Dunbar, P. R., Ogg, G. S., Chen, J., Rust, N., van der Bruggen, P. and Cerundolo, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413–416.

Fitzpatrick ,D. R., Germain -Lee, E. and Valle, D., 1995, Isolation and characterisation of rat and human cDNAs encoding a novel putative peroxisomal enoyl-CoA hydratase. *Genomics*, **27**, 457–466.

Foss, D. L., Baarsch, M. J. and Murtaugh, M. P., 1998, Regulation of hypoxanthine phosphoribosyltransferase, glyceraldehyde-3-phosphate dehydrogenase and beta-actin mRNA expression in porcine immune cells and tissues. *Animal Biotechnology*, **9**, 67–78 .

Frye, R. A., Benz, C. C. and Liu, E., 1989, Detection of amplified oncogenes by differential polymerase chain reaction. *Oncogene*, **4**, 1153–1157.

Geisinger , A., Rodriguez, R., Romero , V. and Wettstein R., 1997, A simple method for screening cDNAs arising from the cloning of RNA differential display bands. *Elsevier Trends Journals Technical Tips Online*, http://tto.trends.com, document T01110.

Gress, T. M., Hoheisel , J. D., Lennon , G. G., Zehetner , G. and Lehrach , H., 1992, Hybridisation fingerprinting of high density cDNA filter arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, **3**, 609–619.

Griffin , G. and Krishna , S., 1998, Cytokines in infectious diseases. *Journal of the Royal College of Physicians, London*, **32**, 195–198.

Groenink , M. and Leegwater , A. C. J., 1996, Isolation of delayed early genes associated with liver regeneration using Clontech PCR-select subtraction technique. *Clontechniques*, **XI**, 23–24.

Guimaraes , M. J., Bazan, J. F., Zlotnik , A., Wiles , M. V., Grimaldi , J. C., Lee, F. and McClanahan , T., 1995b, A new approach to the study of haematopoietic development in the yolk sac and embryoid bodies. *Development*, 121, 3335–3346.

Guimeraes , M. J., Lee, F., Zlotnik , A. and McClanahan , T., 1995a, Differential display by PCR:novel findings and applications. *Nucleic Acids Research*, **23**, 1832–1833.

Gurskaya, N. G., Diatchenko , L., Chenchik , P. D., Siebert , P. D., Khaspekov , G. L., Lukyanov , K. A., Vagner, L. L., Ermolaeva , O. D., Lukyanov, S. A. and Sverdlov , E. D., 1996, Equalising cDNA subtraction based on selective suppression of polymerase chain reaction: Cloning of Jurkat cell transcripts induced by phytohemaglutinin and phorbol 12-Myrystate 13-Acetate. *Analytical Biochemistry*, **240**, 90–97.

Hampson , I. N. and Hampson , L., 1997, CCLS and DROP—subtractive cloning made easy. *Life Science News* (A publication of Amersham Life Science), **23**, 22–24.

Hampson , I. N., Hampson , L. and Dexter , T. M., 1996, Directional random oligonucleotide primed (DROP) global amplification of cDNA: its application to subtractive cDNA cloning. *Nucleic Acids Research*. **24**, 4832–4835.

Hampson , I. N., Pope, L., Cowling , G. J. and Dexter , T. M., 1992, Chemical cross linking subtraction (CCLS): a new method for the generation of subtractive hybridisation probes. *Nucleic Acids Research*, **20**, 2899.

Hara, E., Kato, T., Nakada, S., Sekiya , S. and Oda, K., 1991, Subtractive cDNA cloning using oligo(dT)30-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells. *Nucleic Acids Research*, **19**, 7097–7104.

Hatada, I., Hayashizake, Y., Hirotsune , S., Komatsubara , H. and Mukai, T., 1991, A genomic scanning method for higher organisms using restriction sites as landmarks. *Proceedings of the National Academy of Sciences* (*USA*), **88**, 9523–9527.

Hecht , N., 1998, Molecular mechanisms of male sperm cell differentiation. *Bioessays*, **20**, 555–561.

Hedrick, S., Cohen, D. I., Nielsen , E. A. and Davis, M. E., 1984, Isolation of T cell-specific membrane-associated proteins. *Nature*, **308**, 149–153.

Hertz, R., Seckbach , M., Zakin , M. M. and Bar-Tana, J., 1996, Transcriptional suppression of the transferrin gene by hypolipidemic peroxisome proliferators. *Journal of Biological Chemistry*, **271**, 218–224.

Heuval , J. P. V., Clark, G. C., Kohn , M. C., Tritscher , A. M., Greenlee , W. F., Lucier , G. W. and Bell , D. A., 1994, Dioxin-responsive genes: Examination of dose-response relationships using quantitative reverse transciptase-polymerase chain reaction. *Cancer Research*, **54**, 62–68.

Hillier , L. D., Lennon , G., Becker, M., Bonaldo , M. F., Chiapelli , B., Chissoe , S., Dietrich , N., DuBuque, T., Favello , A., Gish , W., Hawkins , M., Hultman , M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris , M., Parsons, J., Prange, C., Rifkin , L., Rohlfing , T., Schellenberg , K., Soares, M. B., Tan, F., Thierry -Meg, J., Trevaskis , E., Underwood , K., Wohldman , P., Waterston , R., Wilson , R and Marra, M., 1996, Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, **6**, 807–828.

Hubank, M. and Schatz, D. G., 1994, Identifying differences in mRNA expression by representational difference analysis. *Nucleic Acids Research*, **22**, 5640–5648.

Hunter , T., 1991, Cooperation between oncogenes. *Cell*, **64**, 249–270.

Ivanova , N. B. and Belyavsky , A. V., 1995, Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Research*, **23**, 2954–2958.

James , B. D. and Higgins , S. J, 1985, *Nucleic Acid Hybridisation* (Oxford: IRL Press Ltd).

Kas-Deelen , A. M., Harmsen , M. C., de Maar, E. F. and van Son, W. J, 1998, A sensitive method for quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, **5**, 622–626.

Kilty, I. and Vickers, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), **10**, 50–51.

Kleinjan, D.-J. and van Heyningen, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, **7**, 1611–1618.

Ko, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, **18**, 5705–5711.

Lake, B. G., Evans, J. G., Cunninghame, M. E. and Price, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, **101**, 241–248.

Lake, B. G., Evans, J. G., Gray, T. J. B., Korosi, S. A. and North, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guiea pig and marmoset. *Toxicology and Applied Pharmacology*, **99**, 148–160.

Lennard, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, **9**, 60–77.

Levy, S., Todd, S. C. and Maecker, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, **16**, 89–109.

Liang, P. and Pardee, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.

Liang, P., Averboukh, L., Keyomarsi, K., Sager, R. and Pardee, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, **52**, 6966–6968.

Liang, P., Averboukh, L. and Pardee, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, **21**, 3269–3275.

Liang, P., Bauer, D., Averboukh, L., Warthoe, P., Rohrwild, M., Muller, H., Strauss, M. and Pardee, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, **254**, 304–321.

Linskens, M. H., Feng, J., Andrews, W. H., Enlow, B. E., Saati, S. M., Tonkin, L. A., Funk, W. D. and Villeponteau, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, **23**, 3244–3251.

Lisitsyn, N., Lisiitsyn, N. and Wigler, M., 1993, Cloning the differences between two complex genomes. *Science*, **259**, 946–951.

Lohmann, J., Schickle, H. and Bosch, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, **18**, 200–202.

Lunney, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Techology*, **17**, 84–94.

Makowska, J. M., Gibson, G. G. and Bonner, F. W., 1992, Species differences in ciprofibrate-induction of hepaic cytochrome P4504A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, **7**, 183–191.

Maldarelli, F., Xiang, C., Chamoun, G. and Zeichner, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, **53**, 39–51.

Mathieu-Daude, F., Cheng, R., Welsh, J. and McClelland, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, **24**, 1504–1507.

McKenzie, D. and Drake, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), **10**,19–20.

McClelland, M., Mathieu-Daude, F. and Welsh, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, **11**, 242–246.

Mechler, B. and Rabbitts, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, **88**, 29–36.

Meyer, U. A. and Zanger, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269–296.

Mohler, K. M. and Butler, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, **28**, 437–447.

Murphy, L. D., Herzog, C. E., Rudick, J. B., Tito Fojo, A. and Bates, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the mdr-1 gene expression. *Biochemistry*, **29**, 10351–10356.

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., Waterman, M. R., Gotoh, O., Coon, M. J., Estabrook, R. W., Gunsalus, I. C. and Nebert, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, **6**, 1–42.

Nishio , Y., Aiello , L. P. and King , G. L., 1994, Glucose induced genes in bovine aortic smooth muscle cells identified by mRNA differential display. *FASEB Journal*, **8**, 103–106.

O'Neill , M. J. and Sinclair , A. H., 1997, Isolation of rare transcripts by representational difference analysis. *Nucleic Acids Research*, **25**, 2681–2682.

Orton , T. C., Adam , H. K., Bentley , M., Holloway , B. and Tucker, M. J., 1984, Clobuzarit: species differences in the morphological and biochemical response of the liver following chronic administration. *Toxicology and Applied Pharmacology*, **73**, 138–151.

Pelkonen , O., Maenpaa, J., Taavitsainen , P., Rautio, A. and Raunio, H., 1998, Inhibition and Induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica* , **28**, 1203–1253.

Philips , S. M., Bendall , A. J. and Ramshaw , I. A., 1990, Isolation of genes associated with high metastatic potential in rat mammary adenocarcinomas. *Journal of the National Cancer Institute*, **82**, 199–203.

Prashar, Y. and Weissman , S. M., 1996, Analysis of differential gene expression by display of 3'end restriction fragments of cDNAs. *Proceedings of the National Academy of Sciences* (*USA*), **93**, 659–663.

Ragno, S., Estrada, I., Butler, R. and Colston , M. J., 1997, Regulation of macrophage gene expression following invasion by *Mycobacterium tuberculosis*. *Immunology Letters*, **57**, 143–146.

Ramana, K. V. and Kohli , K. K., 1998, Gene regulation of cytochrome P450—an overview. *Indian Journal of Experimental Biology*, **36**, 437–446.

Richard, L., Velasco, P. and Detmar, M., 1998, A simple immunomagnetic protocol for the selective isolation and long-term culture of human dermal microvascular endothelial cells. *Experimental Cell Research*, **240**, 1–6.

Rockett , J. C., Esdaile , D. J. and Gibson , G. G., 1997, Molecular profiling of non-genotoxic hepatocarcinogenesis using differential display reverse transcription-polymerase chain reaction (ddRT-PCR). *European Journal of Drug. Metabolism and Pharmacokinetics*, **22**, 329–333.

Rodricks, J. V. and Turnbull , D., 1987, Inter-species differences in peroxisomes and peroxisome proliferation. *Toxicology and Industrial Health*, **3**, 197–212.

Rogler, G., Hausmann , M., Vogl, D., Aschenbrenner , E., Andus, T., Falk, W., Andreesen , R., Scholmerich , J. and Gross, V., 1998, Isolation and phenotypic characterization of colonic macrophages. *Clinical and Experimental Immunology*, **112**, 205–215.

Rohn , W. M., Lee, Y. J. and Benveniste , E. N., 1996, Regulation of class II MHC expression. *Critical Reviews in Immunology*, **16**, 311–330.

Rudin , C. M. and Thompson , C. B., 1998, B-cell development and maturation. *Seminars in Oncology*, **25**, 435–446.

Sakaguchi, N., Berger, C. N. and Melchers , F., 1986, Isolation of a cDNA copy of an RNA species expressed in murine pre-B cells. *EMBO Journal*, **5**, 2139–2147.

Sambrook , J., Fritsch , E. F. and Maniatis , T., 1989, Gel electrophoresis of DNA. In N. Ford, M. Nolan and M. Fergusen (eds), *Molecular Cloning—A laboratory manual*, 2nd edition (New York: Cold Spring Harbour Laboratory Press), Volume 1, pp. 6–37.

Sargent, T. D. and Dawid, I. B., 1983, Differential gene expression in the gastrula of Xenopus laevis. *Science*, **222**, 135–139.

Schena , M., Shalon , D., Heller , R., Chai , A., Brown., P. O. and Davis , R. W., 1996, Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences* (*USA*), **93**, 10614–10619.

Schneider , C., King , R. M. and Philipson , L., 1988, Genes specifically expressed at growth arrest of mammalian cells. *Cell*, **54**, 787–793.

Schneider -Maunoury , S., Gilardi -Hebenstreit , P. and Charnay , P., 1998, How to build a vertebrate hindbrain. Lessons from genetics. *C R Academy of Science III*, **321**, 819–834.

Semenza , G. L., 1994, Transcriptional regulation of gene expression: mechanisms and pathophysiology. *Human Mutations*, **3**, 180–199.

Sewall , C. H., Bell , D. A., Clark, G. C., Tritscher , A. M., Tully , D. B., Vanden Heuvel , J. and Lucier, G. W., 1995, Induced gene transcription: implications for biomarkers. *Clinical Chemistry*, **41**, 1829–1834.

Singh , N., Agrawal, S. and Rastogi , A. K., 1997, Infectious diseases and immunity: special reference to major histocompatibility complex. *Emerging Infectious Diseases*, **3**, 41–49.

Smith , N. R., Li, A., Aldersley , M., High , A. S., Markham , A. F. and Robinson , P. A., 1997, Rapid determination of the complexity of cDNA bands extracted from DDRT-PCR polyacrylamide gels. *Nucleic Acids* Research, **25** , 3552–3554.

Sompayrac , L., Jane, S., Burn., T. C., Tenen , D. G. and Danna , K. J., 1995, Overcoming limitations of the mRNA differential display technique. *Nucleic Acids Research*, **23**, 4738–4739.

St John , T. P. and Davis , R. W., 1979, Isolation of galactose-inducible DNA sequences from Saccharomyces cerevisiae by differential plaque filter hybridisation. *Cell*, **16**, 443–452.

Sun, Y., Hegamyer , G. and Colburn , N. H., 1994, Molecular cloning of five messenger RNAs differentially expressed in preneoplastic or neoplastic JB6 mouse epidermal cells: one is homologous to human tissue inhibitor of metalloproteinases-3. *Cancer Research*, **54**, 1139–1144.

Sung, Y. J. and Denman, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462–464.

Sutton, G., White, O., Adams, M. and Kerlavage, A., 1995, TIGR Assembler; A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9–19.

Suzuki, Y., Sekiya, T. and Hayashi, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82–84.

Syed, V., Gu, W. and Hecht, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264–273.

Uitterlinden, A. G., Slagboom, P., Knook, D. L. and Vijgl, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742–2746.

Ullman, K. S., Northrop, J. P., Verweij, C. L. and Crabtree, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421–452.

Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B. and Paston, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300–304.

Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484–487.

Voeltz, G. K. and Steitz, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during Xenopus early development. *Molecular and Cell Biology*, **18**, 7537–7545.

Vogelstein, B. and Kinzler, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138–141.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13–14.

Wan, J. S., Sharp, S. J., Poirier, G. M.-C., Wagaman, P. C., Chambers, J., Pyati, J., Hom, Y.-L., Galindo, J. E., Huvar, A., Peterson, P. A., Jackson, M. R. and Erlander, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685–1691.

Walter, J., Belfield, M., Hampson, I. and Read, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS, *Life Science News*, **21**, 13–14.

Wang, Z. and Brown, D. D. 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505–11509.

Wawer, C., Ruggeberg, H., Meyer, G. and Muyzer, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928–4929.

Welsh, J., Chada, K., Dalal, S. S., Cheng, R., Ralph, D. and McClelland, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965–4970.

Wong, H., Anderson, W. D., Cheng, T. and Riabowol, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251–258.

Wong, K. K. and McClelland, M., 1994, Stress-inducible gene of Salmonella typhimurium identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639–643.

Wynford-Thomas, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187–201.

Xhu, D., Chan, W. L., Leung, B. P., Huang, F. P., Wheeler, R., Piedrafita, D., Robinson, J. H. and Liew, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787–794.

Yang, M. and Sytowski, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109–114.

Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. and Sakaki, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207–213.

Zhao, X. J., Newsome, J. T. and Cihlar, R. L., 1998, Up-regulation of two *candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121–129.

Zimmermann, C. R., Orr, W. C., Leclerc, R. F., Barnard, C. and Timberlake, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709–715.

Exhibit D of T. Bedillion Declaration
dated 12/04/03
In USSN 09/937,060

US005807522A

United States Patent [19]

Brown et al.

[11] **Patent Number: 5,807,522**

[45] **Date of Patent: Sep. 15, 1998**

[54] **METHODS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES**

[75] Inventors: **Patrick O. Brown**, Stanford; **Tidhar Dari Shalon**, Atherton, both of Calif.

[73] Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Stanford, Calif.

[21] Appl. No.: **477,809**

[22] Filed: **Jun. 7, 1995**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 261,388, Jun. 17, 1994, abandoned.

[51] **Int. Cl.**^{6} **C12M 1/34; C12M 1/40**

[52] **U.S. Cl.** 422/50; 422/52; 422/55; 422/56; 422/57; 422/68.1; 422/69; 422/82.05; 422/82.06; 422/82.07; 422/82.08; 435/6; 435/7.1; 436/501; 530/300; 530/333; 530/334; 530/350; 536/25.3

[58] **Field of Search** 435/6, 7.1, 172.3; 536/23.1, 24.31, 25.3; 935/78, 3, 19, 80; 436/501, 813; 422/50, 52, 55, 56, 57, 68.1, 69, 82.05, 82.06–82.08; 530/300, 333, 334, 350

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,730,844	5/1973	Gilham et al.	435/6
4,071,315	1/1978	Chateau	436/518
4,486,539	12/1984	Ranki et al.	436/504
4,556,643	12/1985	Paau et al.	435/5
4,563,419	1/1986	Ranki et al.	435/6
4,591,570	5/1986	Chang	436/518
4,670,380	6/1987	Dattagupta	435/6
4,677,054	6/1987	White et al.	435/6
4,683,195	7/1987	Mullis et al.	435/6
4,683,202	7/1987	Mullis	435/91.2
4,716,106	12/1987	Chiswell	435/6

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

721016A2	7/1996	European Pat. Off. .
WO 90/03382	4/1990	WIPO .
WO 92/10588	6/1992	WIPO .
WO 93/22680	11/1993	WIPO .
WO 95/00530	1/1995	WIPO .
WO 95/15970	6/1995	WIPO .
WO 95/21944	8/1995	WIPO .
WO 95/25116	9/1995	WIPO .
WO 96/17958	6/1996	WIPO .

OTHER PUBLICATIONS

Billings et al., "New Techniques for Physical Mapping of the Human Genome," *FASEB*, 5:28–34 (1991).

Chee, et al., "Accessing Genetic Information with High-Density DNA Arrays", *Science*, 274:610–614 (1996).

Drmanac et al., "DNA Sequence Determination by Hybridization: A Strategy for Efficient Large-Scale Sequencing, "*Science*, 260:1649–1652 (1993).

Drmanac et al., "Laboratory Methods: Reliable Hybridization of Oligonucleotides as Short as Six Nucleotides," *DNA and Cell Biology*, 9:527–534 (1990).

Drmanac et al., "Sequencing by Hybridization: Towards an Automated Sequencing of One Million M13 Clones Arrayed on Membranes," *Electrophoresis*, 13:566–573 (1992).

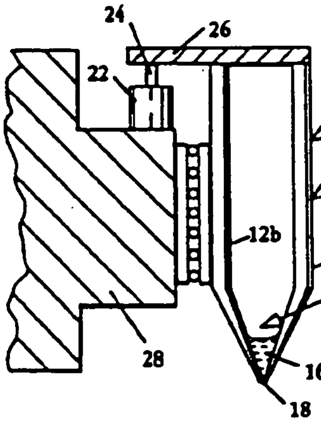
Ekins, et al., "Multianalyte Immunoassay: The Immunological 'Compact Disk'of the Future", *J. Clinical Immunoassay*, 13(4):169–181 (1990).

Primary Examiner—Ardin H. Marschel
Attorney, Agent, or Firm—Arnold White & Durkee

[57] **ABSTRACT**

A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.

7 Claims, 6 Drawing Sheets
(2 of 6 Drawing(s) Filed in Color)



U.S. PATENT DOCUMENTS

4,731,325	3/1988	Palva et al.	435/6
4,755,458	7/1988	Rabbani et al.	435/5
4,767,700	8/1988	Wallace	435/6
4,868,104	9/1989	Kurn et al.	435/6
4,868,105	9/1989	Urdea et al.	435/6
4,921,805	5/1990	Gebeyehu et al.	435/270
4,981,783	1/1991	Augenlicht	435/6
5,013,669	5/1991	Peters, Jr. et al.	436/518
5,028,545	7/1991	Soini	436/501
5,064,754	11/1991	Mills	435/6
5,091,652	2/1992	Mathies et al.	250/458.1
5,100,777	3/1992	Chang	435/7.24
5,143,854	9/1992	Pirrung et al.	436/518
5,185,243	2/1993	Ullman et al.	435/6
5,188,963	2/1993	Stapleton	435/288.3
5,200,051	4/1993	Cozzette et al.	204/403
5,200,312	4/1993	Oprandy	435/5
5,202,231	4/1993	Drmanac et al.	435/6
5,204,268	4/1993	Matsumoto	436/44
5,242,974	9/1993	Holmes	525/54.11
5,252,296	10/1993	Zuckerma et al.	422/116
5,252,743	10/1993	Barrett et al.	548/303.7
5,328,824	7/1994	Ward et al.	435/6
5,338,688	8/1994	Deeg et al.	436/180
5,348,855	9/1994	Dattagupta et al.	435/6
5,389,512	2/1995	Sninsky et al.	435/5
5,412,087	5/1995	McGall et al.	536/24.3
5,434,049	7/1995	Okano et al.	435/6
5,445,934	8/1995	Fodor et al.	435/6
5,472,842	12/1995	Stokke et al.	435/6
5,474,796	12/1995	Brennan	427/2.13
5,474,895	12/1995	Ishii et al.	435/6
5,510,270	4/1996	Fodor et al.	436/518
5,512,430	4/1996	Gong	435/5
5,514,543	5/1996	Grossman et al.	435/6
5,514,785	5/1996	Van Ness et al.	536/22.1
5,516,641	5/1996	Ullman et al.	435/6
5,518,883	5/1996	Soini	435/6
5,545,531	8/1996	Rava et al.	435/6
5,556,748	9/1996	Douglas	435/6
5,556,752	9/1996	Lockhart et al.	435/6
5,563,060	10/1996	Hozier	435/240.23
5,578,832	11/1996	Trulson et al.	250/458.1
5,605,662	2/1997	Heller et al.	422/68.1

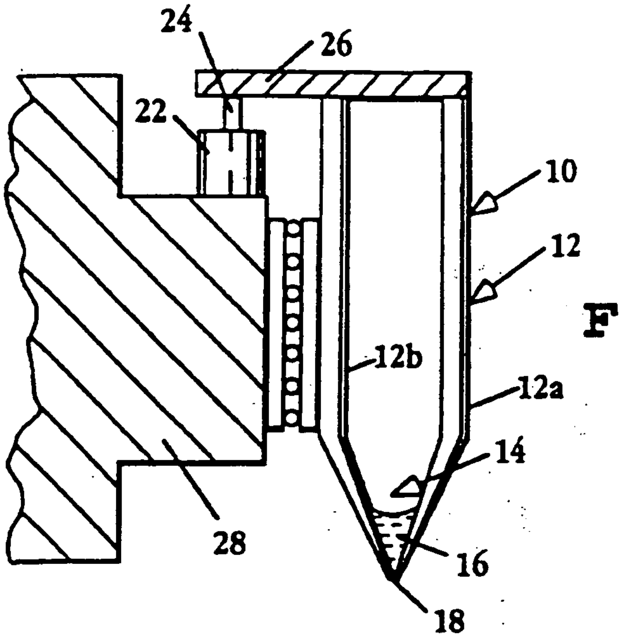


Fig. 1

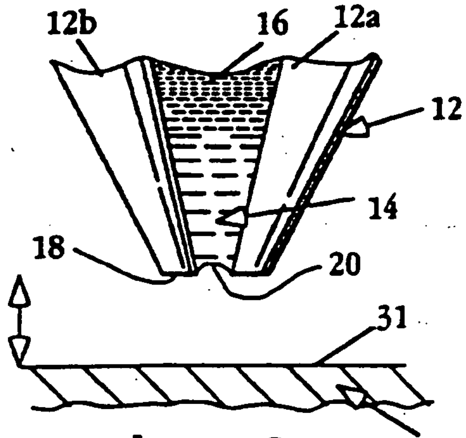


Fig. 2A

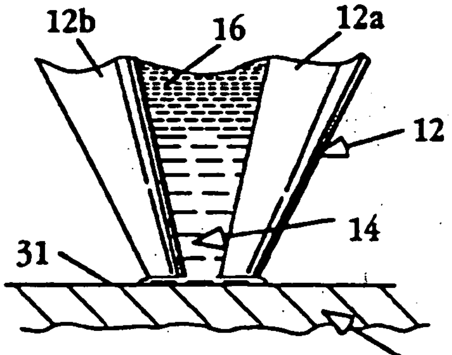


Fig. 2B

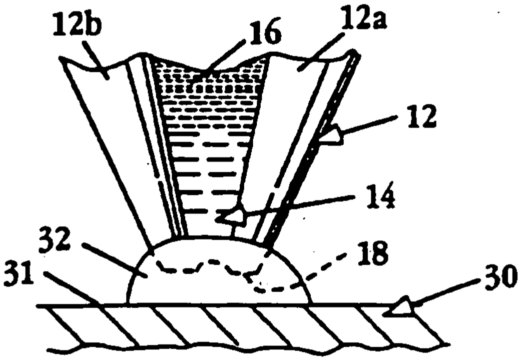


Fig. 2C

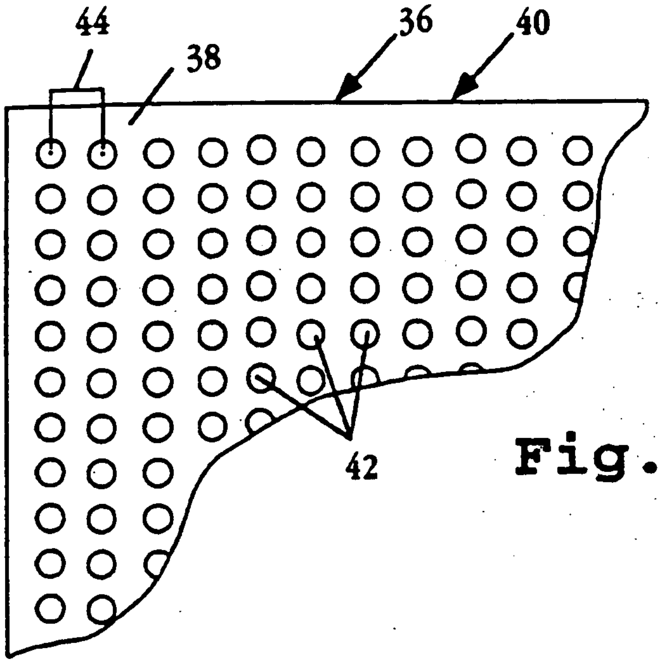

44
38
36
40
42

Fig. 3

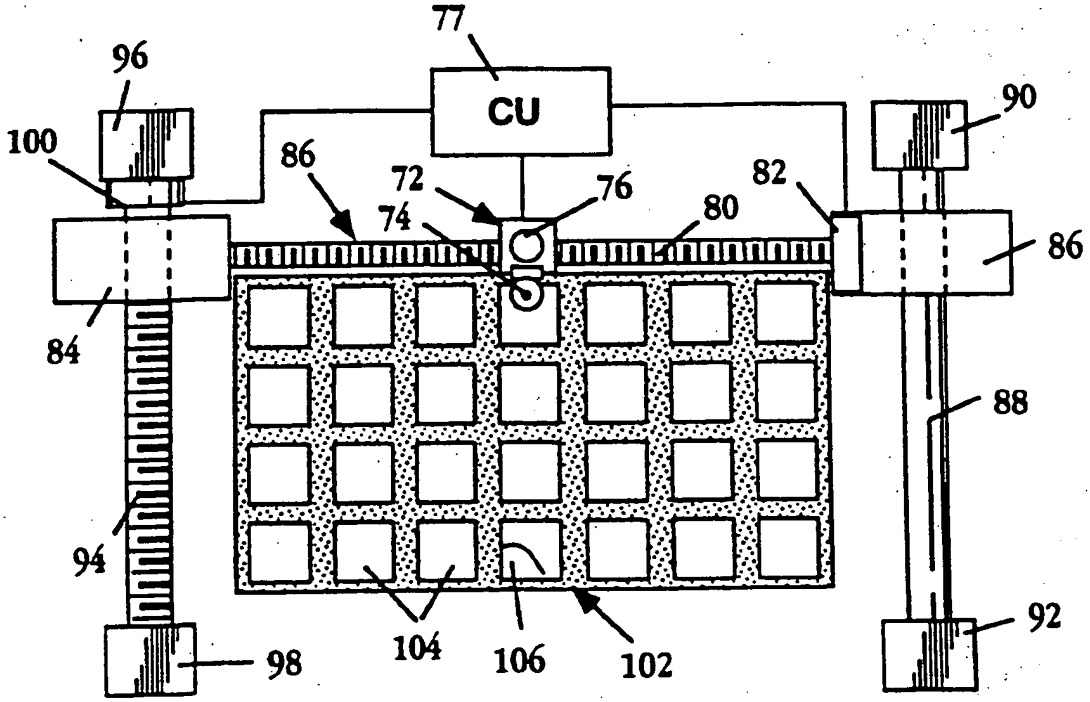

77
CU
96
100
86
72
76
74
80
82
90
86
84
88
94
98
104
106
102
92

Fig. 4

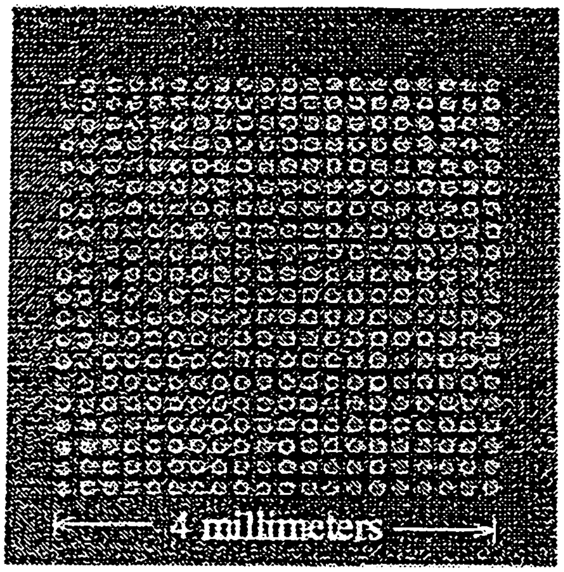


Fig. 5

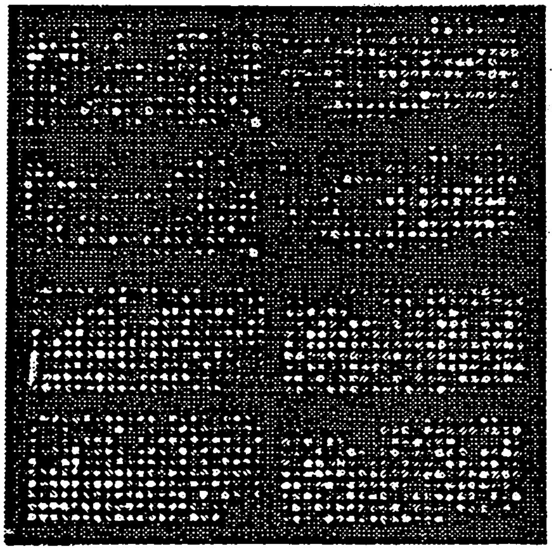


Fig. 6

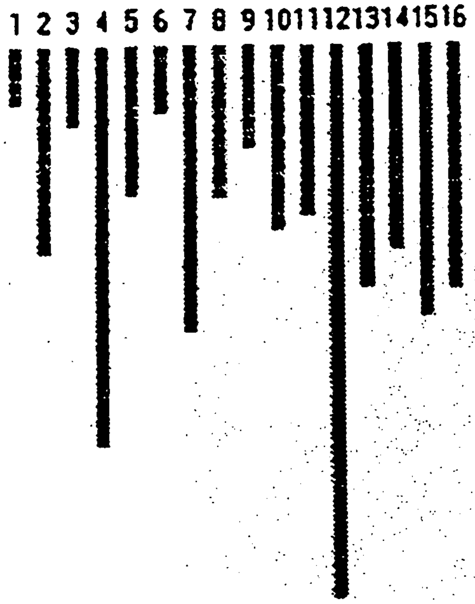


Fig. 7

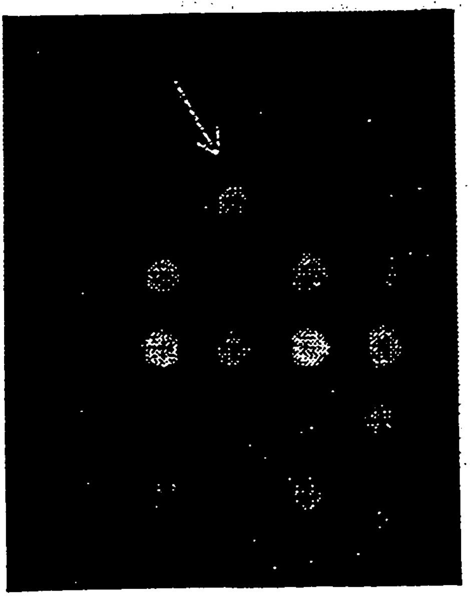


Fig. 8

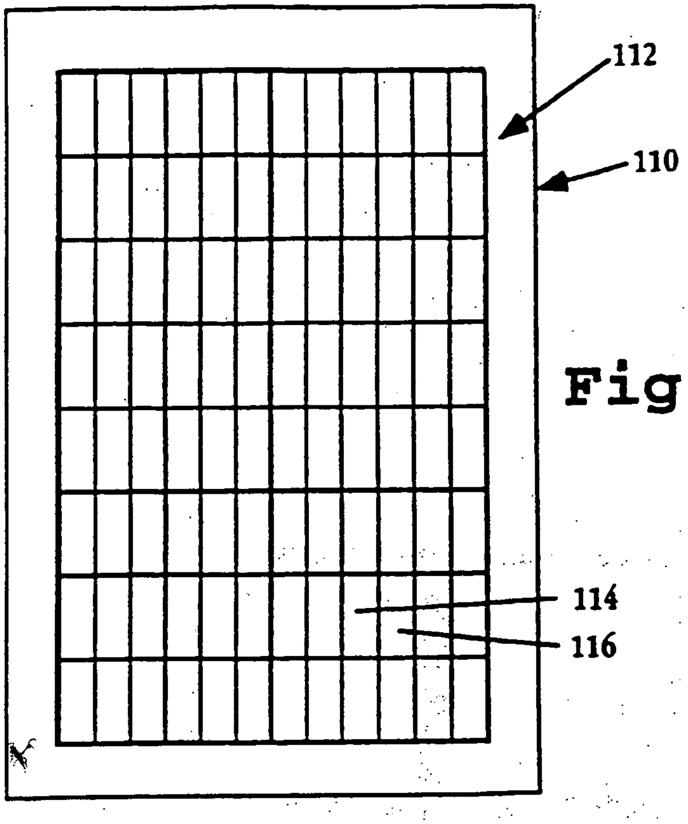


Fig. 9

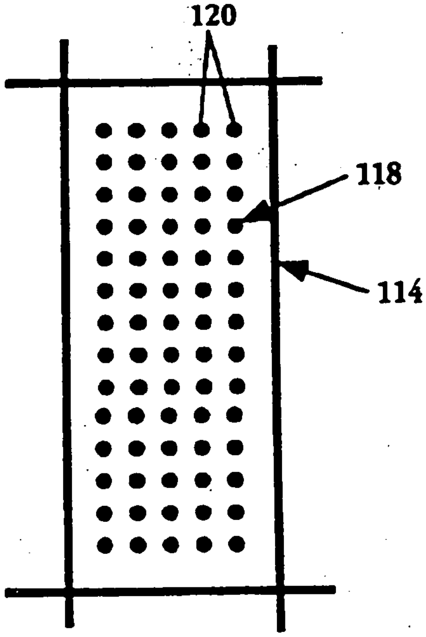


Fig. 10

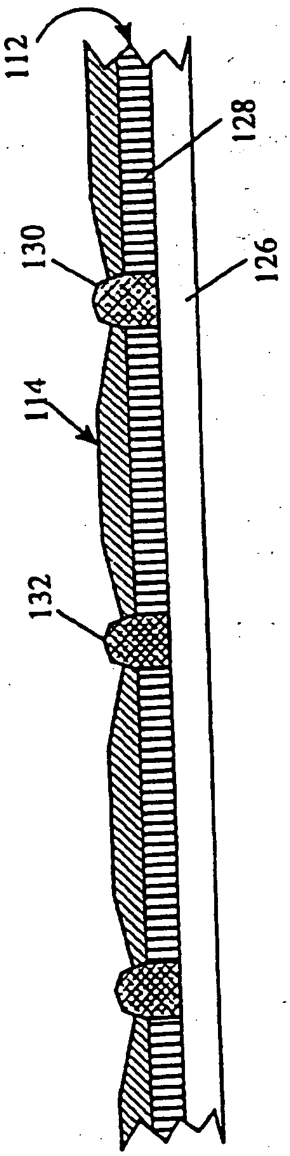


Fig. 11

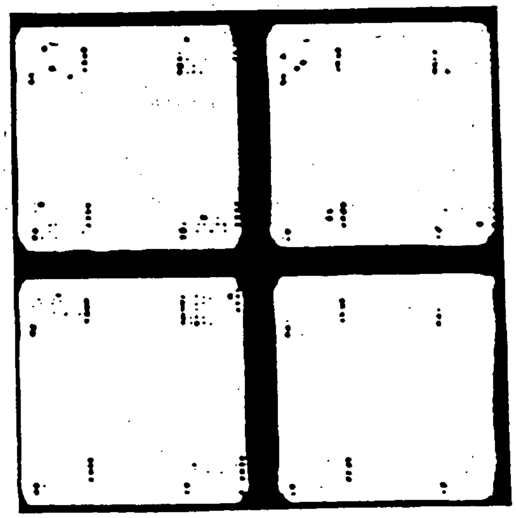


Fig. 12

METHODS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

CROSS-REFERENCE TO RELATED APPLICATION

This application is a continuation-in-part of U.S. patent application Ser. No. 08/261,388, filed Jun. 17, 1994, and now abandoned.

The United States government may have certain rights in the present invention pursuant to Grant No. HG00450 awarded by the National Institutes of Health.

FIELD OF THE INVENTION

This invention relates to a method and apparatus for fabricating microarrays of biological samples for large scale screening assays, such as arrays of DNA samples to be used in DNA hybridization assays for genetic research and diagnostic applications.

REFERENCES

Abouzied, et al., *Journal of AOAC International* 77(2) :495–500 (1994).

Bohlander, et al., *Genomics* 13:1322–1324 (1992).

Drmanac, et al., *Science* 260:1649–1652 (1993).

Fodor, et al., *Science* 251:767–773 (1991).

Khrapko, et al., *DNA Sequence* 1:375–388 (1991).

Kuriyama, et al., *AN ISFET BIOSENSOR, APPLIED BIOSENSORS* (Donald Wise, Ed.), Butterworths, pp. 93–114 (1989).

Lehrach, et al., *HYBRIDIZATION FINGERPRINTING IN GENOME MAPPING AND SEQUENCING, GENOME ANALYSIS*, VOL 1 (Davies and Tilgham, Eds.), Cold Spring Harbor Press, pp. 39–81 (1990).

Maniatis, et al., *MOLECULAR CLONING, A LABORATORY MANUAL*, Cold Spring Harbor Press (1989).

Nelson, et al., *Nature Genetics* 4:11–18 (1993).

Pirrung, et al., U.S. Pat. No. 5,143,854 (1992).

Riles, et al., *Genetics* 134:81–150 (1993).

Schena, M. et al., *Proc. Nat. Acad. Sci. USA* 89:3894–3898 (1992).

Southern, et al., *Genomics* 13:1008–1017 (1992).

BACKGROUND OF THE INVENTION

A variety of methods are currently available for making arrays of biological macromolecules, such as arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation. This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

A more efficient technique employed for making ordered arrays of genomic fragments uses an array f pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22x22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable. In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, et al. (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 50–200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatas, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material. Each vertical stripe is then reacted with a different antigen and the reaction between the immobilized antibody and an antigen is detected using a standard ELISA calorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose. Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen (Research Products International). However, Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the hydrophobic barrier during extended high temperature incubations or in the presence of detergents, which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego Calif.) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate. These plates are capable of containing different reagents in each well without cross-contamination. However, each well is intended to hold only one target element whereas the invention described here makes a microarray of many biomolecules in each subdivided region of the solid support. Furthermore, the 96 well plates are at least 1 cm thick and prevent the use of the device for many calorimetric, fluorescent and radioactive detection formats which require that the membrane lie flat against the detection surface. The invention described here requires no further processing after the assay step since the barriers elements are shallow and do not interfere with the detection step, thereby greatly increasing convenience.

Hyseq Corporation has described a method of making an "array of arrays" on a non-porous solid support for use with their sequencing by hybridization technique. The method described by Hyseq involves modifying the chemistry of the solid support material to form a hydrophobic grid pattern where each subdivided region contains a microarray of biomolecules. Hyseq's flat hydrophobic pattern does not make use of physical blocking as an additional means of preventing cross contamination.

SUMMARY OF THE INVENTION

The invention includes, in one aspect, a method of forming a microarray of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent. The method involves first loading a solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel (i) formed by spaced-apart, coextensive elongate members, (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus. The channel is preferably formed by a pair of spaced-apart tapered elements.

The tip of the dispensing device is tapped against a solid support at a defined position on the support surface with an impulse effective to break the meniscus in the capillary channel, and deposit a selected volume of solution on the surface, preferably a selected volume in the range 0.01 to 100 nl. The two steps are repeated until the desired array is formed.

The method may be practiced in forming a plurality of such arrays, where the solution-depositing step is applied to a selected position on each of a plurality of solid supports at each repeat cycle.

The dispensing device may be loaded with a new solution, by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

Also included in the invention is an automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent. The apparatus has a holder for holding, at known positions, a plurality of planar supports, and a reagent dispensing device of the type described above.

The apparatus further includes a positioning structure for positioning the dispensing device at a selected array position with respect to a support in said holder, and a dispensing structure for moving the dispensing device into tapping engagement against a support with a selected impulse effective to deposit a selected volume on the support, e.g., a selected volume in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are controlled by a control unit in the apparatus. The unit perates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and (iii) dispense the reagent at a defined array position on each of the supports on said holder. The unit may further operate, at the end of a dispensing cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) removing the wash fluid prior to loading the dispensing device with a fresh selected reagent.

The dispensing device in the apparatus may be one of a plurality of such devices which are carried on the arm for dispensing different analyte assay reagents at selected spaced array positions.

In another aspect, the invention includes a substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers in a surface area of less than about 1 cm^2. Each distinct biopolymer (i) is disposed at a separate, defined position in said array, (ii) has a length of at least 50 subunits, and (iii) is present in a defined amount between about 0.1 femtomoles and 100 nanomoles.

In one embodiment, the surface is glass slide surface coated with a polycationic polymer, such as polylysine, and the biopolymers are polynucleotides. In another embodiment, the substrate has a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film. The grid is composed of intersecting water-impervious grid elements extending from said backing to positions raised above the surface of said film, and partitions the film into a plurality of water-impervious cells. A biopolymer array is formed within each well.

More generally, there is provided a substrate for use in detecting binding of labeled polynucleotides to one or more of a plurality different-sequence, immobilized polynucleotides. The substrate includes, in one aspect, a glass support, a coating of a polycationic polymer, such as polylysine, on said surface of the support, and an array of distinct polynucleotides electrostatically bound non-covalently to said coating, where each distinct biopolymer is disposed at a separate, defined position in a surface array of polynucleotides.

In another aspect, the substrate includes a water-impermeable backing, a water-permeable film formed on the backing, and a grid formed on the film, where the grid is composed of intersecting water-impervious grid elements extending from the backing to positions raised above the surface of the film, forming a plurality of cells. A biopolymer array is formed within each cell.

Also forming part of the invention is a method of detecting differential expression of each of a plurality of genes in a first cell type, with respect to expression of the same genes in a second cell type. In practicing the method, there is first produced fluorescent-labeled cDNAs from mRNAs isolated from the two cells types, where the cDNAs from the first and second cell types are labeled with first and second different fluorescent reporters.

A mixture of the labeled cDNAs from the two cell types is added to an array of polynucleotides representing a plurality of known genes derived from the two cell types, under conditions that result in hybridization of the cDNAs to complementary-sequence polynucleotides in the array. The array is then examined by fluorescence under fluorescence excitation conditions in which (i) polynucleotides in the array that are hybridized predominantly to cDNAs derived from one of the first or second cell types give a distinct first or second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to substantially equal numbers of cDNAs derived from the first and second cell types give a distinct combined fluorescence emission color, respectively. The relative expression of known genes in the two cell types can then be determined by the observed fluorescence emission color of each spot.

These and other objects and features of the invention will become more fully apparent when the following detailed description of the invention is read in conjunction with the accompanying figures.

The file of this patent contains at least one drawing executed in color. Copies of this patent with color drawing (s) will be provided by the Patent and Trademark Office upon request and payment of the necessary fee.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a side view of a reagent-dispensing device having a open-capillary dispensing head constructed for use in one embodiment of the invention;

FIGS. 2A–2C illustrate steps in the delivery of a fixed-volume bead on a hydrophobic surface employing the dispensing head from FIG. 1, in accordance with one embodiment of the method of the invention;

FIG. 3 shows a portion of a two-dimensional array of analyte-assay regions constructed according to the method of the invention;

FIG. 4 is a planar view showing components of an automated apparatus for forming arrays in accordance with the invention.

FIG. 5 shows a fluorescent image of an actual 20×20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-l-lysine coated slide, where the total area covered by the 400 element array is 16 square millimeters;

FIG. 6 is a fluorescent image of a 1.8 cm×1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome labeled with a green fluorophore and the other half with a red fluorophore;

FIG. 7 shows the translation of the hybridization image of FIG. 6 into a karyotype of the yeast genome, where the elements of FIG. 6 microarray contain yeast DNA sequences that have been previously physically mapped in the yeast genome;

FIG. 8 shows a fluorescent image of a 0.5 cm×0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild type Arabidopsis plant labeled with a green fluorophore and total cDNA from a transgenic Arabidopsis plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic Arabidopsis plant;

FIG. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

FIG. 10 shows an enlarged plan view of one of the cells in the substrate in FIG. 9, showing an array of polynucleotide regions in the cell;

FIG. 11 is an enlarged sectional view of the substrate in FIG. 9, taken along a section line in that figure; and

FIG. 12 is a scanned image of a 3 cm×3 cm nitrocellulose solid support containing four identical arrays of M13 clones in each of four quadrants, where each quadrant was hybridized simultaneously to a different oligonucleotide using an open face hybridization method.

DETAILED DESCRIPTION OF THE INVENTION

I. Definitions

Unless indicated otherwise, the terms defined below have the following meanings:

"Ligand" refers to one member of a ligand/anti-ligand binding pair. The ligand may be, for example, one of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; an effector molecule in an effector/receptor binding pair; or an antigen in an antigen/antibody or antigen/antibody fragment binding pair.

"Anti-ligand" refers to the opposite member of a ligand/anti-ligand binding pair. The anti-ligand may be the other of the nucleic acid strands in a complementary, hybridized nucleic acid duplex binding pair; the receptor molecule in an effector/receptor binding pair; or an antibody or antibody fragment molecule in antigen/antibody or antigen/antibody fragment binding pair, respectively.

"Analyte" or "analyte molecule" refers to a molecule, typically a macromolecule, such as a polynucleotide or polypeptide, whose presence, amount, and/or identity are to be determined. The analyte is one member of a ligand/anti-ligand pair.

"Analyte-specific assay reagent" refers to a molecule effective to bind specifically to an analyte molecule. The reagent is the opposite member of a ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a linear or two-dimensional array of preferably discrete regions, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about 100/cm^2, and preferably at least about 1000/cm^2. The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10–250 μm, and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

FIG. 1 illustrates, in a partially schematic view, a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below. The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12*a*, 12*b* which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in FIG. 2A. The advantages of the open channel construction of the dispenser are discussed below.

With continued reference to FIG. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in the dispenser on a support, as will be described below with reference to FIGS. 2A–2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring bias, to a normal, raised position, as shown. The dispenser is carried on the piston by a connecting member 26, as shown. The just-described moving structure is also referred to herein as dispensing means for moving the dispenser into engagement with a solid support, for dispensing a known volume of fluid on the support.

The dispensing device just described is carried on an arm 28 that may be moved either linearly or in an x-y plane to position the dispenser at a selected deposition position, as will be described.

FIGS. 2A–2C illustrate the method of depositing a known amount of reagent solution in the just-described dispenser on the surface of a solid support, such as the support indicated at 30. The support is a polymer, glass, or other solid-material support having a surface indicated at 31.

In one general embodiment, the surface is a relatively hydrophilic, i.e., wettable surface, such as a surface having native, bound or covalently attached charged groups. One such surface described below is a glass surface having an absorbed layer of a polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is formed to have a relatively hydrophobic character, i.e., one that causes aqueous medium deposited on the surface to bead. A variety of known hydrophobic polymers, such as polystyrene, polypropylene, or polyethylene have desired hydrophobic properties, as do glass and a variety of lubricant or other hydrophobic films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected analyte-specific reagent solution, such as by dipping the dispenser tip, after washing, into a solution of the reagent, and allowing filling by capillary flow into the dispenser channel. The dispenser is now moved to a selected position with respect to a support surface, placing the dispenser tip directly above the support-surface position at which the reagent is to be deposited. This movement takes place with the dispenser tip in its raised position, as seen in FIG. 2A, where the tip is typically at least several 1–5 mm above the surface f the substrate.

With the dispenser so positioned, solenoid 22 is now activated to cause the dispenser tip to move rapidly toward and away from the substrate surface, making momentary contact with the surface, in effect, tapping the tip of the dispenser against the support surface. The tapping movement of the tip against the surface acts to break the liquid meniscus in the tip channel, bringing the liquid in the tip into contact with the support surface. This, in turn, produces a flowing of the liquid into the capillary space between the tip and the surface, acting to draw liquid out of the dispenser channel, as seen in FIG. 2B.

FIG. 2C shows flow of fluid from the tip onto the support surface, which in this case is a hydrophobic surface. The figure illustrates that liquid continues to flow from the dispenser onto the support surface until it forms a liquid bead 32. At a given bead size, i.e., volume, the tendecy of liquid to flow onto the surface will be balanced by the hydrophobic surface interaction of the bead with the support surface, which acts to limit the total bead area on the surface, and by the surface tension of the droplet, which tends toward a given bead curvature. At this point, a given bead volume will have formed, and continued contact of the dispenser tip with the bead, as the dispenser tip is being withdrawn, will have little or no effect on bead volume.

For liquid-dispensing on a more hydrophilic surface, the liquid will have less of a tendency to bead, and the dispensed volume will be more sensitive to the total dwell time of the dispenser tip in the immediate vicinity of the support surface, e.g., the positions illustrated in FIGS. 2B and 2C.

The desired deposition volume, i.e., bead volume, formed by this method is preferably in the range 2 pl (picoliters) to 2 nl (nanoliters), although volumes as high as 100 nl or more may be dispensed. It will be appreciated that the selected dispensed volume will depend on (i) the "footprint" of the dispenser tip, i.e., the size of the area spanned by the tip, (ii) the hydrophobicity of the support surface, and (iii) the time of contact with and rate of withdrawal of the tip from the support surface. In addition, bead size may be reduced by increasing the viscosity of the medium, effectively reducing the flow time of liquid from the dispenser onto the support surface. The drop size may be further constrained by depositing the drop in a hydrophilic region surrounded by a hydrophobic grid pattern on the support surface.

In a typical embodiment, the dispenser tip is tapped rapidly against the support surface, with a total residence time in contact with the support of less than about 1 msec, and a rate of upward travel from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with the surface is a hemispherical bead, with a diameter approximately equal to the width of the dispenser tip, as shown in FIG. 2C, the volume of the bead formed in relation to dispenser tip width (d) is given in Table 1 below. As seen, the volume of the bead ranges between 2 pl to 2 nl as the width size is increased from about 20 to 200 μm.

TABLE 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface, and/or increasing the viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that the tweezers-like, open-capillary dispenser tip provides the advantages that (i) the open channel of the tip facilitates rapid, efficient washing and drying before reloading the tip with a new reagent, (ii) passive capillary action can load the sample directly from a standard microwell plate while retaining sufficient sample in the open capillary reservoir for the printing of numerous arrays, (iii) open capillaries are less prone to clogging than closed capillaries, and (iv) open capillaries do not require a perfectly faced bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface 38 of a solid support 40 in accordance with the method just described is shown in FIG. 3. The array is formed of a plurality of analyte-specific reagent regions, such as regions 42, where each region may include a different analyte-specific reagent. As indicated above, the diameter of each region is preferably between about 20–200 μm. The spacing between each region and its closest (non-diagonal) neighbor, measured from center-to-center (indicated at 44), is preferably in the range of about 20–400 μm. Thus, for example, an array having a center-to-center spacing of about 250 μm contains about 40 regions/cm or 1,600 regions/cm^2. After formation of the array, the support is treated to evaporate the liquid of the droplet forming each region, to leave a desired array of dried, relatively flat regions. This drying may be done by heating or under vacuum.

In some cases, it is desired to first rehydrate the droplets containing the analyte reagents to allow for more time for adsorption to the solid support. It is also possible to spot out the analyte reagents in a humid environment so that droplets do not dry until the arraying operation is complete.

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in FIG. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to FIG. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in FIGS. 1 and 2A–2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the figure form what is referred to here collectively as a displacement assembly 86.

The displacement assembly is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along an x axis in the figure. In one mode, the assembly functions to move the dispenser in x-axis increments having a selected distance in the range 5–25 μm. In another mode, the dispenser unit may be moved in precise x-axis increments of several microns or more, for positioning the dispenser at associated positions on adjacent supports, as will be described below.

The displacement assembly, in turn, is mounted for movement in the "y" (vertical) axis of the figure, for positioning the dispenser at a selected y axis position. The structure mounting the assembly includes a fixed rod 88 mounted rigidly between a pair of frame bars 90, 92, and a worm screw 94 mounted for rotation between a pair of frame bars 96, 98. The worm screw is driven (rotated) by a stepper motor 100 which operates under the control of unit 77. The motor is mounted on bar 96, as shown.

The structure just described, including worm screw 94 and motor 100, is constructed to produce precise, micro-range movement in the direction of the screw, i.e., along a y axis in the figure. As above, the structure functions in one mode to move the dispenser in y-axis increments having a selected distance in the range 5–250 μm, and in a second mode, to move the dispenser in precise y-axis increments of several microns (μm) or more, for positioning the dispenser at associated positions on adjacent supports.

The displacement assembly and structure for moving this assembly in the y axis are referred to herein collectively as positioning means for positioning the dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of reagent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuation of the dispenser is then effective to dispense a selected-volume droplet of that reagent at this location. As noted above, this operation is effective to dispense a selected volume preferably between 2 pl and 2 nl of the reagent solution.

The dispenser is now moved to the corresponding position at an adjacent support and a similar volume of the solution is dispensed at this position. The process is repeated until the reagent has been dispensed at this preselected corresponding position on each of the supports.

Where it is desired to dispense a single reagent at more than two array positions on a support, the dispenser may be moved to different array positions at each support, before moving the dispenser to a new support, or solution can be dispensed at individual positions on each support, at one selected position, then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is positioned over a wash solution (not shown), and the dispenser tip is dipped in and out of this solution until the reagent solution has been substantially washed from the tip. Solution can be removed from the tip, after each dipping, by vacuum, compressed air spray, sponge, or the like.

The dispenser tip is now dipped in a second reagent well, and the filled tip is moved to a second selected array position in the first support. The process of dispensing reagent at each of the corresponding second-array positions is then carried out as above. This process is repeated until an entire microarray of reagent solutions on each of the supports has been formed.

IV. Microarray Substrate

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.

A. Multi-Cell Substrate

FIG. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8×12 rectangular array 112 of cells, such as cells 114, 116, formed on the substrate surface. With reference to FIG. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in FIG. 3.

The 96-cell array shown in FIG. 9 typically has array dimensions between about 12 and 244 mm in width and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, i.e., between about 1 and 20 in width and 1 and 50 mm in length.

The construction of substrate is shown cross-sectionally in FIG. 11, which is an enlarged sectional view taken along view line 124 in FIG. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed on the surface of the backing is a water-permeable film 128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm. The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to FIG. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In FIG. 11, defined volumes samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of distinct biopolymers. In one general embodiment, the microarrays in the well are identical arrays of distinct biopolymers, e.g., different sequence polynucleotides. Such arrays can be formed in accordance with the methods described in Section II, by depositing a first selected polynucleotide at the same selected microarray position in each of the cells, then depositing a second polynucleotide at a different microarray position in each well, and so on until a complete, identical microarray is formed in each cell.

In a preferred embodiment, each microarray contains about 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2. Also in a preferred embodiment, the biopolymers in each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles. The ability to form high-density arrays of biopolymers, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiment, the biopolymers are polynucleotides having lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by schemes involving parallel, step-wise polymer synthesis on the array surface.

In the case of a polynucleotide array, in an assay procedure, a small volume of the labeled DNA probe mixture in a standard hybridization solution is loaded onto each cell. The solution will spread to cover the entire microarray and stop at the barrier elements. The solid support is then incubated in a humid chamber at the appropriate temperature as required by the assay.

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber. At the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents if needed and analyzed using standard calorimetric, radioactive or fluorescent detection means. All processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of the microarrays on the solid support.

B. Glass-Slide Polynucleotide Array

FIG. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.

The slide is coated by placing a uniform-thickness film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is bound to surface via electrostatic binding between negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-l-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, Mo.).

To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is illustrated in Examples 1 and 2.

To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2. In the embodiment shown in FIG. 5, the microarray contains 400 regions in an area of about 16 mm^2, or 2.5×10^3 regions/cm^2. Also in a preferred embodiment, the polynucleotides in each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-density arrays of this type, where each region is formed of a well-defined amount of deposited material, can be achieved in accordance with the microarray-forming method described in Section II.

Also in a preferred embodiment, the polynucleotides have lengths of at least about 50 bp, i.e., substantially longer than oligonucleotides which can be formed in high-density arrays by various in situ synthesis schemes.

V. Utility

Microarrays of immobilized nucleic acid sequences prepared in accordance with the invention can be used for large scale hybridization assays in numerous genetic applications, including genetic and physical mapping of genomes, monitoring of gene expression, DNA sequencing, genetic diagnosis, genotyping of organisms, and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment is hybridized to an ordered array of DNA fragments, and the identity of the DNA elements applied to the array is unambiguously established by the pixel or pattern of pixels of the array that are detected. One application of such arrays for creating a genetic map is described by Nelson, et al. (1993). In constructing physical maps of the genome, arrays of immobilized cloned DNA fragments are hybridized with other cloned DNA fragments to establish whether the cloned fragments in the probe mixture overlap and are therefore contiguous to the immobilized clones on the array. For example, Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also be used for genetic diagnostics. To illustrate, an array containing multiple forms of a mutated gene or genes can be probed with a labeled mixture of a patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNAs and RNAs can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. An example of this approach is illustrated in Example 2, described with respect to FIG. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cmx2.2 cm microarrays fabricated on a single 12 cmx18 cm sheet of plastic-backed nitrocellulose where each microarray could contain, for example, 100 DNA fragments representing all known mutations of a given gene. The region of interest from each of the DNA samples from 96 patients could be amplified, labeled, and hybridized to the 96 individual arrays with each assay performed in 100 microliters of hybridization solution. The approximately 1 thick silicone rubber barrier elements between individual arrays prevent cross-contamination of the patient samples by sealing the pores of the nitrocellulose and by acting as a physical barrier between each microarray. The solid support containing all 96 microarrays assayed with the 96 patient samples is incubated, rinsed, detected and analyzed as a single sheet of material using standard radioactive, fluorescent, or colorimetric detection means (Maniatas, et al., 1989). Previously, such a procedure would involve the handling, processing and tracking of 96 separate membranes in 96 separate sealed chambers. By processing all 96 arrays as a single sheet of material, significant time and cost savings are possible.

The assay format can be reversed where the patient or organism's DNA is immobilized as the array elements and each array is hybridized with a different mutated allele or genetic marker. The gridded solid support can also be used for parallel non-DNA ELISA assays. Furthermore, the invention allows for the use of all standard detection methods without the need to remove the shallow barrier elements to carry out the detection step.

In addition to the genetic applications listed above, arrays of whole cells, peptides, enzymes, antibodies, antigens, receptors, ligands, phospholipids, polymers, drug cogener preparations or chemical substances can be fabricated by the means described in this invention for large scale screening assays in medical diagnostics, drug discovery, molecular biology, immunology and toxicology.

The multi-cell substrate aspect of the invention allows for the rapid and convenient screening of many DNA probes against many ordered arrays of DNA fragments. This eliminates the need to handle and detect many individual arrays for performing mass screenings for genetic research and diagnostic applications. Numerous microarrays can be fabricated on the same solid support and each microarray reacted with a different DNA probe while the solid support is processed as a single sheet of material.

The following examples illustrate, but in no way are intended to limit, the present invention.

EXAMPLE 1

Genomic-Complexity Hybridization to DNA Microarrays Representing the Yeast *Saccharomyces cerevisiae* Genome with Two-Color Fluorescent Detection

The array elements were randomly amplified PCR (Bohlander, et al., 1992) products using physically mapped lambda clones of *S. cerevisiae* genomic DNA as templates (Riles, et al., 1993). The PCR was performed directly on the lambda phage lysates, resulting in an amplification of both the 35 kb lambda vector and the 5–15 kb yeast insert sequences in the form of a uniform distribution of PCR product between 250–1500 base pairs in length. The PCR product was purified using Sephadex G50 gel filtration (Pharmacia, Piscataway, N.J.) and concentrated by evaporation to dryness at room temperature overnight. Each of the 864 amplified lambda clones was rehydrated in 15 μl of 3xSSC in preparation for spotting onto the glass.

The microarrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine (Sigma). The automated apparatus described in Section III loaded 1 μl of the concentrated lambda clone PCR product in 3xSSC directly from 96 well storage plates into the open capillary printing element and deposited ~5 nl of sample per slide at 380 micron spacing between spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove unabsorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90° for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, Calif.). The six largest chromosomes were isolated in one gel slice and the ten smallest chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, Calif.). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, Ill.) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, Mass.) for the pool containing the six largest chromosomes, and with a fluorescein conjugated nucleotide analog (BMB) for the pool containing ten smallest chromosomes. The two pools were mixed and concentrated using an ultrafiltration device (Amicon, Danvers, Mass.).

Five micrograms of the hybridization probe consisting of both chromosome pools in 7.5 μl of TE was denatured in a boiling water bath and then snap cooled on ice. 2.5 μl of concentrated hybridization solution (5xSSC and 0.1% SDS) was added and all 10 μl transferred to the array surface, covered with a cover slip, placed in a custom-built single-slide humidity chamber and incubated at 60° for 12 hours. The slides were then rinsed at room temperature in 0.1xSSC and 0.1% SDS for 5 minutes, cover slipped and scanned.

A custom built laser fluorescent scanner was used to detect the two-color hybridization signals from the 1.8x1.8

cm array at 20 micron resolution. The scanned image was gridded and analyzed using custom image analysis software. After correcting for optical crosstalk between the fluorophores due to their overlapping emission spectra, the red and green hybridization values for each clone on the array were correlated to the known physical map position of the clone resulting in a computer-generated color karyotype of the yeast genome.

FIG. 6 shows the hybridization pattern of the two chromosome pools. A red signal indicates that the lambda clone on the array surface contains a cloned genomic DNA segment from one of the six largest yeast chromosomes. A green signal indicates that the lambda clone insert comes from one of the ten smallest yeast chromosomes. Orange signals indicate repetitive sequences which cross hybridized to both chromosome pools. Control spots on the array confirm that the hybridization is specific and reproducible.

The physical map locations of the genomic DNA fragments contained in each of the clones used as array elements have been previously determined by Olson and co-workers (Riles, et al.), allowing for the automatic generation of the color karyotype shown in FIG. 7. The color of a chromosomal section on the karyotype corresponds to the color of the array element containing the clone from that section. The black regions of the karyotype represent false negative dark spots on the array (10%) or regions of the genome not covered by the Olson clone library (90%). Note that the six largest chromosomes are mainly red while the ten smallest chromosomes are mainly green, thus matching the original CHEF gel isolation of the hybridization probe. Areas of the red chromosomes containing green spots and vice-versa are probably due to spurious sample tracking errors in the formation of the original library and in the amplification and spotting procedures.

The yeast genome arrays have also been probed with individual clones or pools of clones that are fluorescently labeled for physical mapping purposes. The hybridization signals of these clones to the array were translated into positions on the physical map of the yeast genome.

EXAMPLE 2

Total cDNA Hybridized to Micro Arrays of cDNA Clones with Two-Color Fluorescent Detection

Twenty-four clones containing cDNA inserts from the plant Arabidopsis were amplified using PCR. Salt was added to the purified PCR products to a final concentration of 3xSSC. The cDNA clones were spotted on poly-l-lysine coated microscope slides in a manner similar to Example 1. Among the cDNA clones was a clone representing a transcription factor HAT4, which had previously been used to create a transgenic line of the plant Arabidopsis, in which this gene is present at ten times the level found in wild-type Arabidopsis (Schena, et al., 1992).

Total poly-A mRNA from wild type Arabidopsis was isolated using standard methods (Maniatis, et al., 1989) and reverse transcribed into total cDNA, using a fluorescein nucleotide analog to label the cDNA product (green fluorescence). A similar procedure was performed with the transgenic line of Arabidopsis where the transcription factor HAT4 was inserted into the genome using standard gene transfer protocols. cDNA copies of mRNA from the transgenic plant are labeled with a lissamine nucleotide analog (red fluorescence). Two micrograms of the cDNA products from each type of plant were pooled together and hybridized to the cDNA clone array in a 10 microliter hybridization reaction in a manner similar to Example 1. Rinsing and detection of hybridization was also performed in a manner similar to Example 1. FIG. 8 shows the resulting hybridization pattern of the array.

Genes equally expressed in wild type and the transgenic Arabidopsis appeared yellow due to equal contributions of the green and red fluorescence to the final signal. The dots are different intensities of yellow indicating various levels of gene expression. The cDNA clone representing the transcription factor HAT4, expressed in the transgenic line of Arabidopsis but not detectably expressed in wild type Arabidopsis, appears as a red dot (with the arrow pointing to it), indicating the preferential expression of the transcription factor in the red-labeled transgenic Arabidopsis and the relative lack of expression of the transcription factor in the green-labeled wild type Arabidopsis.

An advantage of the microarray hybridization format for gene expression studies is the high partial concentration of each cDNA species achievable in the 10 microliter hybridization reaction. This high partial concentration allows for detection of rare transcripts without the need for PCR amplification of the hybridization probe which may bias the true genetic representation of each discrete cDNA species.

Gene expression studies such as these can be used for genomics research to discover which genes are expressed in which cell types, disease states, development states or environmental conditions. Gene expression studies can also be used for diagnosis of disease by empirically correlating gene expression patterns to disease states.

EXAMPLE 3

Multiplexed Colorimetric Hybridization on a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was gridded with barrier elements made from silicone rubber according to the description in Section IV-A. The sheet was soaked in 10xSSC and allowed to dry. As shown in FIG. 12, 192 M13 clones, each with a different yeast inserts were arrayed 400 microns apart in four quadrants of the solid support using the automated device described in Section III. The bottom left quadrant served as a negative control for hybridization, while each of the other three quadrants was hybridized simultaneously with a different oligonucleotide using the open-face hybridization technology described in Section IV-A. The first two and last four elements of each array are positive controls for the calorimetric detection step.

The oligonucleotides were labeled with fluorescein, which was detected using an anti-fluorescein antibody conjugated to alkaline phosphatase that precipitated an NBT/BCIP dye on the solid support (Amersham). Perfect matches between the labeled oligos and the M13 clones resulted in dark spots visible to the naked eye and detected using an optical scanner (HP ScanJet II) attached to a personal computer. The hybridization patterns are different in every quadrant indicating that each oligo found several unique M13 clones from among the 192 with a perfect sequence match. Note that the open capillary printing tip leaves detectable dimples on the nitrocellulose which can be used to automatically align and analyze the images.

Although the invention has been described with respect to specific embodiments and methods, it will be clear that various changes and modification may be made without departing from the invention.

We claim:

1. A method of forming a microarray of discrete analyte-assay regions on a solid support, where each discrete region in the microarray has a selected, analyte-specific reagent, said method comprising,

(a) loading an aqueous solution of a selected analyte-specific reagent in a reagent-dispensing device having an elongate capillary channel adapted to hold a quantity of the reagent solution and having a tip region at which the solution in the channel forms a meniscus,

(b) tapping the tip of the dispensing device against a solid support at a defined position on the surface, with an impulse effective to break the meniscus in the capillary channel and deposit a selected volume between 0.002 and 2 nl of solution on the surface, and

(c) repeating steps (a) and (b) until said microarray is formed.

2. The method of claim 1, wherein the reagents used to form the discrete regions in the microarray are distinct nucleic acid strands and wherein steps (a) and (b) are repeated until the microarray has about 100 or more discrete regions of distinct nucleic acid strands per cm^2 of solid support.

3. The method of claim 1, wherein the reagents used to form the discrete regions in the microarray are distinct nucleic acid strands and wherein steps (a) and (b) are repeated until the microarray has about 1000 or more discrete regions of distinct nucleic acid strands per cm^2 of solid support.

4. The method of claim 2, wherein the channel is open-sided.

5. The method of claim 3, wherein the channel is open-sided.

6. The method of claim 4, wherein the volume is between 0.002 and 0.25 nl.

7. The method of claim 5, wherein the volume is between 0.002 and 0.25 nl.

* * * * *